KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering Science

Vertical Transistors

A slippery path towards the ultimate CMOS scaling

Dmitry Yakimets

Supervisors:
Prof. dr. ir. Kristin De Meyer
Dr. ir. Nadine Collaert

Dissertation presented in partial fulfilment of the requirements for the degree of Doctor in Engineering Science: Electrical Engineering

January 2017

Vertical Transistors

A slippery path towards the ultimate CMOS scaling

Dmitry YAKIMETS

Examination committee:
Prof. dr. Adhemar Bultheel, chair
Prof. dr. ir. Kristin De Meyer, supervisor
Dr. ir. Nadine Collaert, supervisor
Prof. dr. ir. Georges Gielen
Prof. dr. ir. Wim Dehaene
Prof. dr. Lars-Erik Wernersson
(Lund University)
Dr. Praveen Raghavan
(imec)

Dissertation presented in partial fulfilment of the requirements for the degree of Doctor in Engineering Science: Electrical Engineering

January 2017

Acknowledgments

This dissertation is the result of a four-year research. Four years is quite a long time. Therefore, it is not a big surprise that a lot of people influenced my work. I would like to thank all of them for their contributions to my professional and/or personal growth.

First, I would like to express my gratitude to my supervisors Prof. Kristin De Meyer and Dr. Nadine Collaert. I thank them for the provided research freedom combined with a delicate guidance, and for the continuous support throughout my work. In addition, I would like to acknowledge all of my examination committee members for their valuable feedback.

Next, a special thanks goes to Aaron Thean, who is now a professor at the National University of Singapore. It is his vision, which largely inspired me to conduct my research. I guess I should also thank him for the given possibility to work towards my PhD at imec. My direct managers, either on the department level (Morin Dehan, Praveen Raghavan) or on the INSITE program level (Karim Mercha and Alessio Spessot), also contributed to my research by a careful balance of my exposure to the imec partners. They quite managed to protect me from the daily routine requests and to burden with truly interesting questions.

Talking about imec industrial partners, they played a very important role in steering of my PhD. I was privileged to get a direct feedback and insight from the leading companies around the globe which tremendously improved the quality of my work. The list of people I want to thank here is long. I will focus on assignees working at imec. To name a few, I thank Mustafa Badaroglu (Qualcomm), Victor Moroz (Synopsys), Gerben Dornboos (TSMC), Aryan Afzalian (TSMC), Bartek Pawlak (GlobalFoundries).

I am very thankful to my past and present imec colleagues. They have always been patient answering my numerous questions and have created a very nice atmosphere: Doyoung Jang, Pieter Schuddinck, Trong Huynh Bao, Yasser Sherazi, Peter Debacker, Arindam Mallik, Bharani Chava, Sushil Sakhare,

Kenichi Miyaguchi, Geert Eneman, Anne Verhulst, Tarun Agarwal Kumar. The list of course by far not complete and I want to apologize upfront if I have not included you here. Also, I want to specially acknowledge my daily advisor, Marie Garcia Bardon, for sharing her experience with me. I wish this experience was enriching for her at least a bit of what it was for me.

The quite large Russian-speaking community in Leuven has always represented a piece of home for me: Pavel Poliakov, Andriy Hikavyy, Alexey Milenin, Mikhail Krishtab, Marina Timmermans, Vladimir Leonov, Ivan Zyulkov, Nina Tutunjyan, Alexander Mityashin, Sergey Osokin, Pavlo Fesenko, Aleksandra Zharikova, Alexey Kubarev, Davit Melkonyan, Stella Deheryan... Again, I am sure that some of the names are missing here. In case you are in the missing list, please do not take it personally.

Last, but not least, I want to thank my family for their unwavering love, encouragement and endless support. A very special thanks is for my wife, Anna Yakimets, along with my son, Lev Yakimets, for being on my side whatever happens and for the everyday inspiration.

Abstract

The semiconductor industry has largely relied on Moore's law, based on the observation that every new generation of transistors has been better than the previous one in Power, Performance, Area and Cost (PPAC) metrics simultaneously. However, this trend is under a pressure now. The main issue is related to the enormous complexity of both technology and design, which drastically raises not only the manufacturing, but also the R&D costs. Therefore, in order to minimize risks and maximize benefits of a new technology, it is being co-optimized hand in hand with a design relying on this technology.

The scaling of lateral transistors is going to reach its limit soon because it mainly relies on the scaling of contacted gate pitch (CGP), which, in turn, forces the scaling of gate length, S/D spacers and contacts. Reduction of any of these dimensions is undesirable as it leads to poorer electrostatic control, increased parasitic capacitance and increased access resistance, respectively. There are lateral devices, like nanowire-based FETs, which may postpone the problem of CGP budgeting but they cannot solve it.

The focus of this PhD work is on the vertical devices. These devices are less constrained on gate length and spacer thickness as they are oriented vertically and thus should demonstrate better scalability than lateral transistors. We quantify the advantages of the vertical devices in terms of PPA metrics through a holistic benchmark by combining the design techniques and technology limitations which are likely to be in place at the 5nm technology. In order to do this, we perform the layouts analysis, model and evaluate RC parasitics, calibrate compact models to TCAD and experimental data. Afterwards, we run simulations on a ring oscillator level to extract the PPA metrics.

We have not limited ourselves to the conventional MOSFETs only, but we also benchmark vertical III-V heterojunction Tunnel FETs in order to get a better understanding under which conditions the vertical architecture is the most advantageous. This allows us to shed light on the ultimate CMOS scaling and to understand whether the introduction of vertical transistors can enable the next technological nodes.

Beknopte samenvatting

Vooruitgang in de halfgeleiderindustrie heeft lange tijd vertrouwd op de wet van Moore, die gebaseerd is op de observatie dat elke volgende generatie van transistoren beter bleek te zijn tegenover de vorige m.b.t. vermogen, performantie, oppervlakte en kostprijs (PPAC). Recentelijk is deze trend onder steeds meer druk te komen staan. Het belangrijkste probleemaspect hierin is de enorme complexiteit van zowel de technologie als het ontwerp; hetgeen niet alleen de productiekost dramatisch doet toenemen, maar ook de kosten gerelateerd aan onderzoek en ontwikkeling (R&D). Om de risico's te beperken en om de voordelen van elke nieuwe technologie te maximaliseren, is het noodzakelijk geworden om de nieuwe technologie simultaan te optimaliseren met zijn bijhorende circuitontwerpen.

Het voortdurend schalen (verkleinen) van laterale transistoren zal binnenkort zijn limiet bereiken omdat die sterk vertrouwt op de schaling van het kanaalinterval, die op zijn beurt een schaling afdwingt van de kanaallengte, S/D spaties en contacten. Gelijk welke daling in elk van deze dimensies, is zeer ongewenst, omdat dit aanleiding geeft tot slechtere elektrostatische controle, grotere parasitaire capaciteiten en ingangsweerstanden. Het probleem van de kanaalintervalschaling kan enigszins uitgesteld worden door gebruik te maken van o.a. nanodraad gebaseerde veldeffectschakelaars, maar een echte oplossing bieden deze transistoren niet.

De focus van dit doctoraatswerk ligt op de verticale transistoren. Deze schakelaars zijn minder beperkt in kanaallengte en spatiedikte, omdat zij verticaal gericht zijn, en bijgevolg zouden zij een betere schaalbaarheid dan laterale transistoren moeten bezitten. We kwantificeren de voordelen van verticale schakelaars aan de hand van de eerder genoemde PPA-maatstaven, door middel van een allesomvattende testomgeving met een combinatie van ontwerptechnieken en technologische beperkingen die men kan verwachten voor de 5nm technologie. Om ons doel te bereiken, analyseren we eerst de opmaak (layout) van verschillende schakelingen, modeleren en evalueren we de bijhorende parasitaire weerstanden en capaciteiten, kalibreren we ons transistormodel aan de hand van computersimulaties (TCAD) en meetdata. Nadien voeren we

ringoscillatorsimulaties uit om de PPA-maatstaven te kunnen extraheren.

Hierbij hebben we ons niet beperkt tot de conventionele metaal-oxide-halfgeleider veldeffectschakelaars (MOSFET), maar hebben dus ook verticale III-V heterojunctie tunnel FETs onderzocht om aldus een beter inzicht te krijgen onder welke voorwaarden de verticale architectuur het meest aangewezen zou zijn. Dit onderzoek laat ons toe om licht te werpen op de ultieme CMOS-schaling en om bijgevolg ook te begrijpen of het invoeren van verticale transistoren de stap naar de volgende technologie toelaat.

Contents

List of Abbreviations

BEOL back end of line. 47, 48, 63, 96, 101, 111–113
BR ballistic ratio. 20, 56–59
BTBT band-to-band tunneling. 73

CGP contacted gate pitch. 2–7, 44, 63, 64, 66, 67, 90, 92, 97, 102, 109–111
CM compact model. 9

DIBL drain-induced barrier lowering. 58, 76
DOS density of states. 73
DSP digital signal processor. 63
DTCO design-technology co-optimization. 1
DVFS dynamic voltage-frequency scaling. 91, 93

EDA electronic design automation. 11
EDP energy-delay product. 93
EOT equivalent oxide thickness. 47
EUV extreme ultraviolet. 3, 4, 46

FEOL front end of line. 63, 111
FPGA field-programmable gate array. 49, 111

GAA gate all-around. 6
GIDL gate-induced drain leakage. 18
GPU graphics processing unit. 63

HDL hardware description language. 12

IC integrated circuit. 11–13, 15, 28, 36
IoT Internet of things. 112

LDPC low-density parity-check. 63, 64

MOL middle of line. 111

MOSFET metal-oxide-semiconductor field-effect transistor. 8–10, 17–19, 47, 48, 71–75, 81, 83, 86, 87, 89, 96, 98, 100, 102, 104, 106, 107, 110–112
MP metal pitch. 2–4, 43, 90, 109, 110

NM noise margin. 81–87, 104, 106, 107, 112
NSh nanosheet. 47–49, 52–55, 57–59, 61, 62, 65, 66, 90
NW nanowire. 6, 23, 28, 29, 33, 41, 44, 45, 47–49, 52–55, 57, 58, 60–62, 65, 66, 68, 69, 90, 91, 94, 97, 98, 101, 102, 104, 111, 113
NWFET nanowire FET. 6, 101, 110

PEX parasitic extraction. 16, 33
PPAC Power-Performance-Area-Cost. 8, 9, 44, 47, 109, 112
PSG phosphorus silicate glass. 7, 91

RO ring oscillator. 93, 99, 111
ROI return on investment. 39
RTL register transfer level. 12

SADP self-aligned double patterning. 4
SAQP self-aligned quadrupole patterning. 40, 46
SCE short channel effect. 5
SoC system on chip. 63, 68, 91, 102, 109, 111
SPICE simulation program with integrated circuit emphasis. 12
SRAM static random access memory. 81, 83, 110, 113
SS subthreshold slope. 58, 60, 71, 76, 80, 82, 83, 92, 93, 104, 105, 111

TCAD technology computer-aided design. 13–15, 19, 21, 50, 109, 110
TFET tunnel field-effect transistor. 8–10, 16, 35, 72–75, 77–87, 89, 96, 98–100, 102, 104–107, 110–113
TLM transmission line model. 22–25

VDP vertical device pitch. 44, 45
VFET vertical MOSFET. 8, 10, 29, 30, 44, 45, 48, 49, 68, 74, 78, 79, 84, 89, 92, 98, 102, 104, 110–112
VTC voltage transfer characteristics. 81–83, 86, 87

List of Symbols

α	activity factor
τ	switching delay
C	capacitance
$C_{Channel}$	channel capacitance
C_{GD}	gate to drain capacitance
C_{GS}	gate to source capacitance
C_{ox}	oxide capacitance
D	diameter
E_F	Fermi energy
$E_{G,eff}$	effective bandgap
E_C	conduction band
E_V	valence band
$FinH$	fin height
$FinW$	fin width
H_l	cell height of lateral architecture
H_v	cell height of vertical architecture
N	cell's width in a number of stacked devices
I	current
I_{DS}	drain to source current
I_{Dlin}	drain current in linear regime
I_{Dsat}	drain current in saturation regime
I_{OFF}	leakage current
k_B	Boltzmann constant, $k_B = 8.62 \times 10^{-5}$ eV/K
L_G	gate length
N_{sp}	extra source pocket doping in tfets
N_d	drain doping
N_p	pocket doping in tfets
N_s	source doping
P_{active}	active power

$P_{leakage}$	leakage power
P_{NW}	pitch between the nanowires
P_{total}	total power
q	elementary charge, $q = 1.60 \times 10^{-19}$ C
R	resistance
R_{SD}	s/d total access resistance
R_D	drain resistance
R_S	source resistance
T	temperature
T_{SP}	S/D spacer thickness
T_C	S/D contact size
V	voltage
V_{Dlin}	linear drain voltage
V_{Dsat}	saturation drain voltage
V_{DD}	positive supply voltage
V_{DS}	drain to source bias
V_{GS}	gate to source bias
V_{IN}	input voltage
V_{OUT}	output voltage
V_{SS}	ground reference voltage
V_{TH}	threshold voltage
W_{EFF}	effective width
α_μ	mobility degradation factor
μ_{app}	apparent mobility
$\mu_{long\ channel}$	long channel mobility
μ_e	effective mobility
ε	electrical field
ε_c	critical electrical field
I_{bal}	ballistic current
N_{inv}	number of inversion carriers
v	carrier velocity
$R_{ch.bal}$	channel resistance in ballistic limit
$R_{ch.scat}$	channel resistance with scattering in the channel
Q_i	inversion charge sheet density
v_{inj}	injection velocity
v_{sat}	saturation velocity
v_d	average carrier velocity near the drain
v_s	average carrier velocity near the source
ϵ	dielectric permittivity
C_{elec}	capacitance between two perpendicular electrodes
C_{ext}	capacitance to extention
C_{side}	capacitance between the sides of electrodes

C_{PP}	parallel plate capacitance
k	elliptic integral modulus
k'	complementary modulus of the elliptic integral, $k' = \sqrt{1-k^2}$
L_{un}	length of undoped part
ρ_{sc}	semiconductor resistivity
ρ_c	specific contact resistivity
ρ_m	metal resistivity
L_T	transfer length
R_c	contact resistance

CHAPTER 1

Introduction

1.1 Logic scaling landscape

1.1.1 Overview

Advancements in microelectronics have been driven by innovations on various levels: from the introduction of new materials up to the usage of the multicore processors architecture. On top the device and design innovations, transistors have scaled simultaneously with an increase in number of transistors per chip with every new technology. Modern designs consist of billions of transistors which are linked together through an extraordinary complex interconnect network of metal wires. In order to be printed, these wires have to comply with various lithography constraints. Similarly, a device engineer should take into account a number of process limitations to design a transistor, *e.g.* the minimum gate length is governed by the gate stack thickness. That said, technology limitations should be accounted for in order to design a chip which would not only meet the specs but would also be manufacturable. It works the other way to: a technology is going to be useful only in case it is possible to make a design which meets the performance targets. In the following introductory sections we will go a little deeper into the scaling challenges to highlight the necessity of the design-technology co-optimization (DTCO).

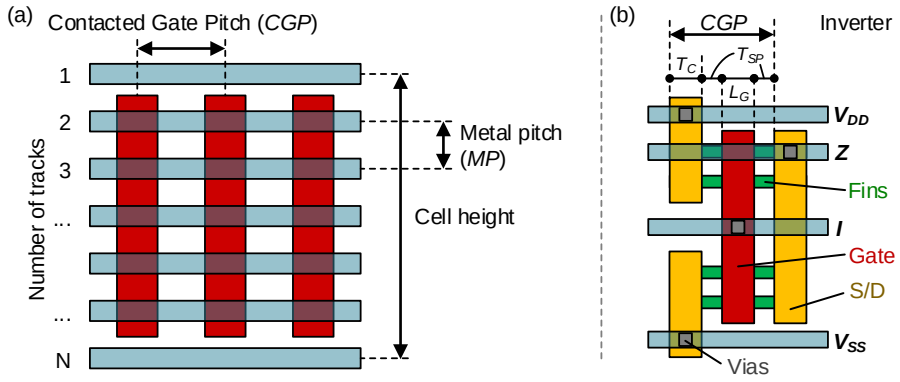


Figure 1.1: a) Technology ground rules allow to quickly estimate a footprint of a standard cell. b) Sketch of an inverter layout made with FinFETs. The definition of contacted gate pitch is illustrated.

1.1.2 What is a technology node?

The first challenge is to give a definition to a technology. At the time this text is written, Intel is about to launch its 10 nm technology node in mass production. Its main competitors, like Samsung and TSMC, start to disclose details on their 10 nm technology and plans for further scaling [1], [2]. However, not all the 10 nm technology nodes are the same.

For a while, the node name was corresponding to either the first layer metal half-pitch or the gate length. These days, the node name mainly serves the marketing purpose. Still, people keep on talking about the Moore's law [3], mostly about its second revision: the number of transistors per chip doubles every two years thanks to a higher density of components per chip, *i.e.* smaller transistors [4]. Two years has been the cadence of the new technology nodes introduction. Linking these two inputs, we may claim that transition from one node to another node happens once the area of the same design scales by roughly a factor of two. To enable area comparison of various technologies without doing actual designs, the so-called ground rules are used. These rules are the contacted gate pitch (CGP), the bottom layer metal pitch (MP) and the number of metal tracks needed to finish the logic cell (Fig. 1.1a).

Let us first outline the scaling landscape by looking at the recent ground rules scaling trends (Fig. 1.2). Going backwards in history, we limit ourselves to the Intel 22 nm node, the first node featuring FinFETs. We may see that the ground rules chosen by different companies are different even within the same node name. Intel has followed the conventional 0.5× area scaling quite precisely,

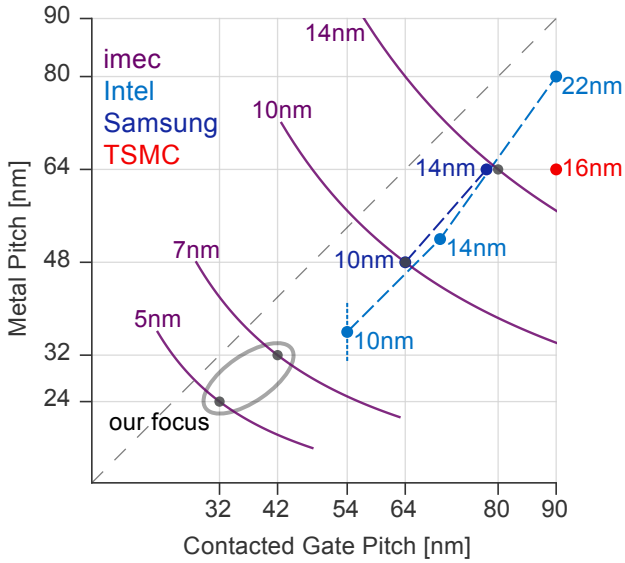


Figure 1.2: Technological nodes having same name do not necessarily have the same ground rules. However, the transition from one node to another comes with roughly 0.5× scaling of MP and CGP product.

while Samsung scaled their 14 nm to 10 nm in MP-CGP product metric by "just" 0.6×.

Yet, there is another rule: the number of metal tracks. Samsung does not disclose whether they reduced their standard cell heights, but Intel does. Intel did it to achieve more than 0.5× scaling [5]. The reason behind is related to the economics — the core of the Moore's law. To keep on scaling the cost per function from one node to another, these functions have to be implemented in the aggressively scaled silicon footprint to compensate for the dramatic increase of the wafer cost for the latest technologies [6]. The large wafer cost mainly originates from the very expensive lithography processes.

1.1.3 Lithography

The lithography has become a dominant component of the wafer cost because in order to print tiny layout features multi-patterning techniques are required. This, in turn, results in a large number of masks. Extreme ultraviolet (EUV) lithography might help to reduce the number of masks (by about 25% [6]).

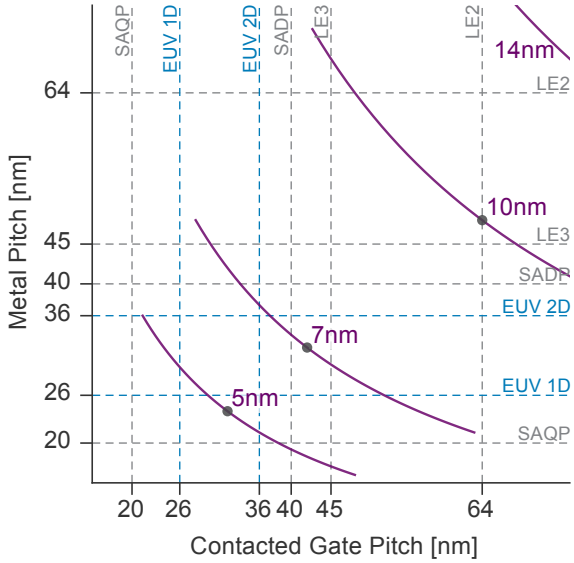


Figure 1.3: Various patterning cliffs should be taken into account when ground rules are being defined. Blue colors correspond to the single exposure limits of the EUV lithography. Grey colors are for the limits of various multi-patterning techniques in 193i lithography. Grey points indicate the nominal ground rules for imec technologies.

However, it is still not adopted by the industry. Moreover, given the pace of scaling, even EUV lithography will have hard time patterning features at the 5 nm node like dimensions. The so-called patterning cliffs (Fig. 1.3), dimensions beyond which certain lithography techniques do not work, have to be taken into account when technology ground rules are defined [7]. The iso-area curves for imec nodes in Fig. 1.2 indicate that various combinations of MP and CGP may be chosen for a certain technology. This choice should be guided by *both* electrical and patterning considerations. Cost-wise it is better to operate close to the cliff to get as much profit as possible from a given lithography option. This is the reason why at 7 nm node we chose CGP of 42 nm — very close to the standard 193 nm immersion ArF-based lithography self-aligned double patterning (SADP) cliff [7].

Multi-patterning solutions put constraints on design, *e.g.* the bottom most metal layers in the latest technologies are laid out with uni-directional shapes. These metal lines should be on the grid coupled together with the underlying device dimensions like CGP [7], [8]. This has an impact on the overall interconnect

scheme as signal routing might get quite complex due to lack of design freedom.

1.1.4 Interconnects

Not only the routing gets more complex, but also the interconnect RC-delay increases. This is largely related to the fact that thickness of metal wires becomes small with scaling which results in drastic increase of copper resistivity and, thus, wire resistance [9]. To overcome this issue, the state-of-art software tools which physically distribute the wiring between the devices tend to quickly propagate signals to the upper metal layers. Unfortunately, this propagation also does not come for free, as there is quite some voltage drop on vias in between the layers.

Because of this non-negligible interconnect RC-delay induced either by wires themselves or by vias in between the metal layers, it is crucial to include these wires into performance analysis of a particular technology option, which is often neglected in literature.

1.1.5 Device

Despite an increasing role of interconnects, it is still a device, what defines a technology performance. A classical lateral device consists of three regions (Fig. 1.1b): gate defined by the gate length (L_G), spacers between gate and source / drain (S/D) regions (T_{SP}), and S/D contacts (T_C). All of these regions scale with CGP scaling as CGP $= L_G + 2T_{SP} + T_C$. A gate length scaling comes with a problem of electrostatics control [10]. A spacer thickness scaling might cause reliability issues and causes a major increase in parasitic capacitance between gate and S/D contacts as this capacitance is roughly proportional to $1/T_{SP}$ [11]. A contact size scaling results in an increase of access resistance.

Channel control

Planar devices were replaced by FinFETs because the latter provides much better control over the channel wrapping it from three sides (Fig. 1.4a) [10]. However, to maintain control over short channel effects (SCEs), a fin thickness should be scaled together with a gate length, which may result in too strong variability [12]. On top of that, an increasing role of surface roughness might result in poor drive currents for ultra narrow fins [13], [14]. Moreover, there are some concerns regarding the mechanical stability of tall, high aspect ratio fins [15]. Lateral gate all-around (GAA) FETs (Fig. 1.4b) made of nanowires

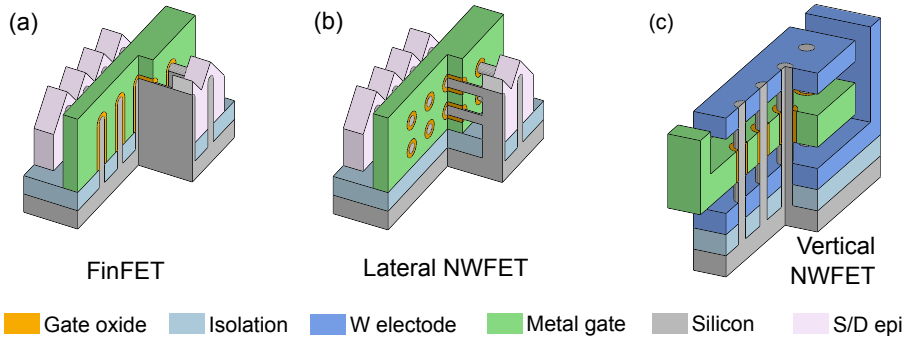


Figure 1.4: 3D sketches of different devices: a) FinFET, b) lateral NWFET, and c) vertical NWFET.

(NWs) allow relaxed channel dimensions with respect to a fin thicknesses while providing similar SCE control [16]. This suggests their better scalability and puts off the problem of gate length scaling but does not solve it. The best solution would be to make the gate length independent on CGP. This might be achieved by rotating the channel perpendicular to the wafer surface — by using the vertical nanowire FET (NWFET) (Fig. 1.4c).

Parasitics

As we focus on the advanced technologies in this thesis, let us omit the discussion on RC parasitics for planar devices and concentrate on FinFETs. Why should we discuss device parasitics? This is because parasitic capacitances have already exceed intrinsic channel capacitance values [17]. Similar thing might happen with resistance: channel resistance lowers thanks to various innovation related to the intrinsic transport enhancement (*e.g.* mechanical stress introduction). Access resistance, however, increases and might soon exceed channel resistance [17]. The two dominant resistive components are the contact resistance and the extension resistance (region underneath the S/D spacers) [18].

Contact resistance is defined by the contacting area and by the specific contact resistivity between a metal and a semiconductor. Both may be carefully engineered. For example, depending on the fin height and the fin pitch one can either epitaxially grow diamond shaped features in the S/D regions or place a contact directly on top of the fins [19], [20]. The specific contact resistivity mainly depends on the doping level in the semiconductor and the interface quality. Theoretical studies [21] show that there is still some room for

resistivity improvement with respect to the state of art [22], [23]. Nevertheless, the contacting area will inevitably decrease due to smaller contact size T_C.

As for the extension resistance, it is defined by the cross-section area of the extension, the spacer length (T_{SP}) and the doping level in this extension. With scaling, fins get taller, thinner, and positioned at tighter fin pitch [24]. Therefore, the ion implantation techniques get more and more tricky to use [25]. Other solutions, like doping outer diffusion from, *e.g.*, phosphorus silicate glass (PSG) gain interest [26]. Yet, even with this technique, it is hard to achieve high doping levels in the extensions together with steep junctions at the interface with the channel. The next knob to scale an extension resistance is a reduction of a spacer thickness. However, T_{SP} should be scaled carefully not to penalize reliability and not to increase parasitic capacitance too much.

The dominant parasitic capacitances in the modern technologies are the fringing capacitances between the gate electrode and S/D extensions and capacitances between the gate and S/D electrodes. The best way to reduce them is to lower the permittivity of the spacer material. However, there are integration and reliability challenges associated with these materials. The state of art permittivity value for the spacer material is around 4.4 [27], although there is work being done on air gaps integration [28], which is an ultimate improvement possible. However, even with the air gaps, distance between the gate and S/D electrodes will keep on shrinking with scaling due to CGP reduction. Again, it would be highly beneficial for further scaling to have spacers independent on CGP. Vertical architecture makes T_{SP} and T_C independent from the CGP and thus looks very attractive for the most advanced technological nodes. However, there is very little work done on device assessment which would account for the limitations coming from these advanced technologies and would benchmark vertical and lateral devices [29], [30], [31], [32], [33].

1.1.6 Design-Technology Co-Optimization

For the holistic device exploration, either vertical or lateral, various design and technology aspects should be considered. The design and technology should be co-optimized in order to find the next node device which would be possible to manufacture, would provide sufficient drive to meet application performance requirements, would not consume too much energy per switch, and would be cost efficient. Typically, there requirements are referred to as Power-Performance-Area-Cost (PPAC) metrics. Optimization of these metrics is a global effort and this work is a part of it focusing on understanding if a transistor with *vertically* oriented channels might be such a future device.

Power is of a particular interest as more and more applications are related to mobile, battery limited cases. As power is proportional to the square of V_{DD}, its scaling would be highly efficient for power reduction. In order to scale V_{DD} and to maintain reasonable device performance, a certain overdrive ($V_{DD} - V_{TH}$) should be preserved. V_{TH} scaling is problematic with conventional metal-oxide-semiconductor field-effect transistors (MOSFETs), therefore the steep slope devices gain a lot of momentum as they allow V_{TH} scaling without jeopardizing leakage. Tunnel field-effect transistors (TFETs) are, probably, the most mature devices belonging to the family of the steep slope devices [34]. The best TFETs are made of hetero-junction III-V materials [35]. Due to their complex structure, they are hard to manufacture laterally [36]. Therefore, it was natural to include them into our analysis of vertical devices. This work, thus, is not about a vertical MOSFET, but about the benefits and drawbacks of vertical architecture in general with a focus on a vertical device.

1.2 Objectives of the work

Vertical MOSFETs (VFETs) (Fig. 1.4c) are less constrained on the gate length and the spacer thickness as they are oriented vertically and thus should demonstrate better scalability than lateral devices. It does not mean, however, that the gate length should be relaxed largely with respect to the lateral devices. In the case of too long gates, the device drive current degrades together with an increase in the channel capacitance. Similar thinking might be applied to S/D spacers. The extension resistance would become too high limiting the device performance if spacers are too long.

Vertical devices are not new and have been proposed quite some time ago [37]. This work, however, investigates the behavior of vertical devices at *scaled* technologies in a holistic way. We do not limit ourselves to the conventional MOSFETs, but go beyond and also look into TFETs, trying to understand whether vertical architecture in general makes sense at scaled dimensions.

What does make vertical architecture so different from the lateral one? Layouts of VFETs-based standard cells are different from the layouts made with lateral devices. They are difficult to interpret, because from the top-down view all the electrodes (top, gate, and bottom) overlap. The device is inherently asymmetric with the source not being equal to the drain because the connection to the bottom electrode should be done through the deep and narrow (and, thus, quite resistive) via. The device optimizations are different, because different constraints apply to vertical devices. The original research presented in this

dissertation focuses on deep understanding of these differences and on their impact on the PPAC metrics. The main objectives of the work are the following:

- Determine the dimensions of vertical device considering the process limitations and fulfilling the requirements on area scaling which is expected from the 5 nm technological node.
- End up with the compact model of a vertical MOSFET, which would be applicable to the determined device dimensions in a sense that it should capture the quasi-ballistic nature of carriers transport and account for quantum confinement effects.
- Formulate analytically all the relevant vertical device parasitics and implement them in a compact model form by describing them in the Verilog-A language.
- Set up the compact model of a vertical TFET, which would capture essential device behaviour and would be easily calibrated against reference TCAD data.
- Holistically optimize lateral device for the 5 nm node like dimensions in order to conduct a fair benchmark with vertical devices at iso-area.
- Benchmark all the above-mentioned device options (vertical MOSFET and TFET, lateral devices) and identify their possible applications.
- Deliver understanding of the observed benchmarking results.

The novelty of this work is related to the fact that we have conducted the benchmarking holistically by combining the design techniques and technology limitations which are likely to be in place at the 5 nm node. This allowed us to shed light on the ultimate CMOS scaling and to understand whether the introduction of vertically oriented channels is sufficient to enable next technological nodes. The results of this work are presented in six chapters.

1.3 Outline of the dissertation

Chapter 2 gives an understanding on how we built a compact model (CM) of a vertical device. First, this chapter identifies the role of a compact model (CM) in the overall chip design flow. We share our vision on how the good CM should look like and briefly discuss how CMs may be implemented. Afterwards, there is a section which describes the limitations of the nowadays industry-standard MOSFET compact model. This section also includes our proposal on how this model might be extended to capture the quasi-ballistic transport in the highly confined channels. Last, we introduce analytical models of VFET-related parasitic resistance and capacitances. We have used these models for the device benchmarking afterwards.

Chapter 3 is related to vertical architecture. In here, we talk about standard cell layouts and discuss what differences vertical device brings with respect to a lateral device in terms of layout efficiency. On top that we quantify the impact of device asymmetry (source is not equal to drain), and we estimate cell parasitics, which are sometimes referred to as middle-of-line parasitics.

Chapter 4 is fully devoted to the vertical conventional MOSFET made under the 5 nm design rules. We look into the DC performance as well as the ring oscillator level performance of various vertical MOSFETs. Specifically, we benchmark transistors with nanosheet-like channels and with nanowire-like channels of different dimensions looking into the detailed comparison of their intrinsic performance and related parasitics.

Chapter 5 is dedicated to the vertical TFET. We introduce the TFET device and a compact model which describes this device. Afterwards, we do the detailed analysis of different metrics similar to the previous chapter. In here, we also spend quite some time discussing the impact of the delayed onset in output characteristics, which is typical for TFETs, on the circuit performance.

In the beginning of **Chapter 6** we propose the methodology on how to benchmark lateral and vertical devices. The device optimization exercise is done afterwards for the lateral FinFETs and NWFETs in order to fix the reference comparison point. Afterwards, the results from the previous two chapters are combined with the results obtained for the optimized lateral devices. They form the comprehensive benchmark of lateral and vertical architectures on the ring oscillator level.

Finally, **Chapter 7** summarizes the findings from all the other chapters and provides general conclusions regarding the applicability of vertical architecture. This summary is extended with the proposals for the possible future research.

CHAPTER 2

Building a Compact Model

2.1 What is a compact model?

2.1.1 IC design implementation flow

The semiconductor industry has followed Moore's law remarkably well, which resulted in enormously complex integrated circuits (ICs) consisting of billions of transistors. Manual placing and wiring of all these components is just impossible. The electronic design automation (EDA) is here to help human with the design. Given the size of the problem, it has to be tackled at various abstraction levels and it is typically done by various tools. Together, these tools form a design flow which may start from a technological process simulation and finish with a thermal analysis of a printed circuit board with lots of chips on it.

For this dissertation, the upper abstraction limit is a single digital IC design. Moreover, the focus of this work is on a single transistor, but we will further demonstrate that understanding of a chip design is crucial for a successful transistor evaluation. Figure 2.1 shows a typical simplified implementation flow of a digital IC.

First, a standard cell library should be prepared. It consists of various electronic logic gates such as inverter, NAND or flip-flop. These gates are presented through different views like physical layout or electrical schematics. There are ways to generate basic layouts automatically, but normally it is a full custom design where all the cells are made manually. This process is constrained by various design rules which come from lithography assumptions, technological

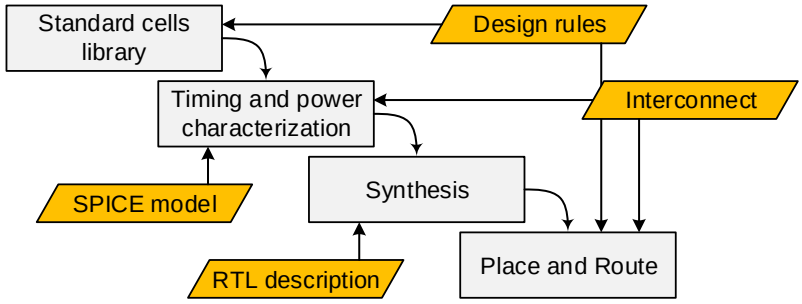


Figure 2.1: IC design implementation flow.

limitations, *etc.* With every new technological node the number of design rules increases which reflects the increasing process complexity.

Once a library is complete, all of its cells should be characterized for timing and power. Typically, there are hundreds of standard cells in a library, which makes the characterization process quite computationally demanding as all the possible input combinations and transitions between them should be analyzed. In order to accurately predict electrical behaviour of the cells, good models of transistors, related parasitic resistances and capacitances (parasitics) and cell interconnects (Fig. 2.2) are necessary.

Transistors and other circuit components are modeled with compact models, which are sometimes called SPICE models (SPICE stands for the simulation program with integrated circuit emphasis). The name reflects its purpose: a compact model is simple, based on analytical equations, and quick to run to minimize a computational burden. Yet, it captures the essential device behaviour. Transistor related parasitics are typically also attributed to a compact model. Cell interconnects are defined through parasitics of metal wiring inside the standard cells. Information on wiring is extracted from cells' layouts. Afterwards, it is converted to the *RC* network based on the interconnect models linking wires geometry and material properties.

Next, it is necessary to map the complex logic function of an IC onto basic cells existing in a standard cell library. A hardware description language (HDL) code is used to describe an IC at a behavioural level. It provides a higher level of abstraction than schematics or layouts. This code is also called the register transfer level (RTL) description of an IC. A synthesis tool maps HDL code onto a library minimizing area while meeting timing and/or power constraints by a

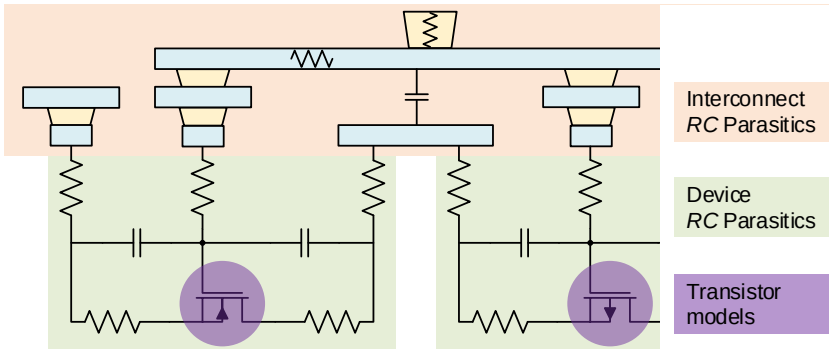


Figure 2.2: Modeling layers.

careful selection of the most suited standard cells. The outcome of a synthesis is a netlist which contains information on all the used cells together with the information on connectivity between them.

Finally, this netlist should be converted into a physical IC layout. This procedure consists of two steps: place and route. Placing allocates a certain position to each cell on the IC layout and routing establishes the interconnects between all of these cells. This process is also based on various design rules. The finalized design is characterized for power, performance and area including all the required interconnects (signoff step).

The accuracy of this characterization is thus mainly driven by the accuracy of compact models. In case the technological process is mature enough, compact models are calibrated on measurements. However, this flow cannot be applied in the case of exploratory (pre-silicon) research due to the lack of measurement data. This brings a question on how to be sure that a compact model has a predictive power towards advanced technologies.

2.1.2 Predictive compact model

For exploratory research, compact models are fitted to the data obtained with technology computer-aided design (TCAD) tools. These data may include simulations of technological process steps along with the device performance predictions. Process simulations are essential to develop and optimize a technology. Although some technological information, like channel doping or stress levels, may also be used by compact models (it depends on a compact model, though). The device structure obtained with process simulations is

normally used as an input for the device simulations. These simulations result in the electrical characteristics of semiconductor devices, which serve as the basis for the further compact models calibration (*e.g.* for the *I-V* and *C-V* fitting).

Figure 2.3 lists typical transport models used in semiconductor devices TCAD-based simulations starting from the least accurate but fastest. The choice of a particular model depends on the size of the problem. In general, the smaller the device is, the more advanced transport model should be used. Figure 2.4 provides rough guidelines on which model has to be used when. The drift diffusion model has worked fine for a long time, being calibrated to hardware or Monte Carlo simulations. However, in order to accurately *predict* the behaviour of the advanced devices, models which are more complex the the drift-diffusion model are required, and their complexity increases exponentially with scaling.

In case there are difficulties in the nanoscale transport description with finite element modelling (TCAD is essentially based on finite elements), it is wrong to expect the analytical compact models to capture transport precisely. Yet, compact models for various devices are being published regularly with an intention to capture more and more physics in order to enhance their predictive capabilities. The major flaw in this trend is related to the loss of one of the main compact model feature: simplicity. On top of that, the majority of the developed compact models are truly predictive only in a limited range of dimensions and are

	Model	Improvements
Semi-classical approaches	Drift-diffusion equations	Good for devices down to 0.5 µm
	Hydrodynamic equations	Velocity overshoot is captured properly
	Boltzmann transport equation	Accurate up to the classical limits
Quantum approaches	Quantum hydrodynamics	All hydrodynamic features + quantum corrections
	Quantum Monte-Carlo methods	All classical features + quantum corrections
	Quantum-Kinetic equation	Accurate up to single particle description
	Green's function methods	Includes correlations in both space and time domain
	Schrödinger equation	Can be solved only for a small number of particles

Figure 2.3: Transport models sorted by complexity: most computationally intensive and accurate are at the bottom. Adapted from [38].

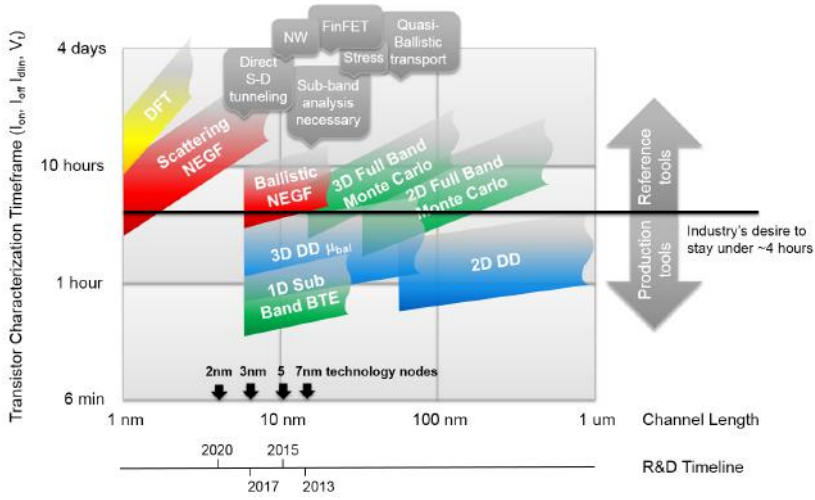


Figure 2.4: Transport models introduction timeline [after V. Moroz, Synopsys].

rarely well scalable because the physics of carriers transport changes with scaling (Fig. 2.4). In that sense, compact models which qualitatively capture the correct device behaviour at all the possible device terminals voltage combinations, provide good convergence and have a limited (but sufficient to be flexible) set of calibration parameters appear as most interesting for exploratory work.

2.1.3 Device parasitics modeling

Device parasitic resistances and capacitances impact circuit performance by reducing operating frequency and increasing energy consumption. Modeling of these components (Fig. 2.2) is significantly easier than that of the transport in the channel as both resistances and capacitances are passive components. Therefore, the development of a *predictive* compact model is justified as it is likely going to be well scalable. In addition, the verification of parasitics compact model is easier because of relatively simple physics.

There are attempts to rely on the mixed-mode TCAD for circuit simulations [39]. In this case, an intrinsic carriers transport, device and cell parasitics are all modeled with TCAD. Unfortunately, this approach is not much viable for a complex IC design as it takes too much time. Such a mixed-mode setup can be used for a compact model verification.

Quite obviously, device *RC*-parasitics depend on its geometry. There are special

tools which extract device-related parameters (*e.g.* electrode dimensions, or channel thickness) from layouts. Typically, on top of the device recognition job, these tools also extract the parasitics present in the interconnect network. This is a step performed during the standard cell characterization and it is called parasitic extraction (PEX). The tool which performs this extraction is called parasitics extractor. An extractor does its job very fast but it has insufficient accuracy and flexibility when the *device* parasitics are concerned. For example, it would be very hard to extract the metal-semiconductor interface resistance in a way it will be described in one of the next sections of this chapter.

Therefore, there are typically two linked problems to solve when an interface between a parasitic compact model and an extractor is designed. 1. Where a particular parasitic (R, C or both) should be computed — in the model or within the PEX flow? 2. How to formulate device description in a way that an extractor can extract these description in an automated way?

The first question will be briefly discussed further on. The second one implies the need of a good documentation of a parasitics model and a coordination between a parasitics model developer with a person who manages PEX. We find this a bit beyond the scope of the thesis and, thus, we have omitted this part.

2.1.4 Verilog-A for compact models

Early compact models were written in C or FORTRAN languages and were designed to work with particular simulators. It quickly became clear that decoupling compact models from simulators brings significant advantages for both developers and users [40]. Next step was to move the development to an unified and standard hardware description languages such as Verilog-A. A compact model written in Verilog-A requires just one tenth of lines of C code and it is free from any simulator interface code. These days most commercial simulators support Verilog-A compact models and compile them so efficiently that the runtime becomes comparable to C coded models [41].

The main drawback of a Verilog-A model is that it still needs to be compiled prior to run and this compilation is taking quite some time in case of complex models. Therefore, the most important compact models which are widely used by the industry are still written in C and incorporated into simulators. A number of emerging devices, such as TFETs, do not have any industry accepted compact models (Compact Model Coalition is a representative of the industry here [42]). Thus, the exploratory work is nearly exclusively done in Verilog-A which has became the *de facto* language for compact models [43]. There will be a description of the developed analytical model of vertical device parasitics

further in the text. This model has also been programmed in Verilog-A to support circuit simulations.

2.2 Quasi-ballistic transport in MOSFET

2.2.1 Limits of drift-diffusion model

The main compact model for the multi-gate MOSFETs is BSIM-CMG [10]. The beauty of this model is that it has proved to be robust and flexible, it supports a number of secondary physical effects (*e.g.* self-heating, noise, *etc.*), and it operates with parameters which reflect device structure very well. However, BSIM-CMG is essentially a drift-diffusion model, which means that with scaling, it cannot capture the effects related to the increase of ballisticity in the transport [44].

The drift-diffusion model fails to predict both saturation (high drain bias) and linear (low drain bias) currents [45], [46]. An explanation comes from the fact that the drift-diffusion theory is based on a *low field simplification* of the Boltzmann Transport Equation [47]. These low field near- equilibrium conditions are only found in a long channel MOSFETs. For nanoscale devices, because of the steep junctions and small feature sizes, there are large built-in electric fields even at low drain bias. Therefore, the low field assumption breaks and the drift-diffusion model looses its predictive power.

In order to correctly predict the current in long channel devices at high drain bias, the drift-diffusion model has been extended with a concept of saturation velocity. It is an *empirical* feature introducing the dependency of mobility on electric field under and above saturation velocity [48], [49]. There are several ways to express this dependency. The one in Eq 2.1 is probably the most known [48].

$$v = \frac{\mu_e \varepsilon}{[1 + (\varepsilon/\varepsilon_c)^n]^{(1/n)}}, \tag{2.1}$$

where n (≥ 1) defines how quickly the carriers reach saturation velocity, ε_c is the critical field. At low fields, the carriers velocity v is proportional to the field ε through the effective mobility μ_e. However, at large fields, $v = v_{sat} = \mu_e \varepsilon_c$.

Although the saturation velocity has helped to capture transport behavior at high fields, it still does not work well for short channel devices because they exhibit a rapid spatial variation of potential. In such cases, some fraction of the carriers may acquire much higher than thermal energy (typically, near the

drain). In this case the velocity overshoot may happen: the carrier velocity exceeds the saturation velocity [50].

In this sense, despite the fact that the drift-diffusion model has been used for a long time already, it is more of an empirical model rather than physical model, which means that over time it has been extended with various fitting parameters to make it very flexible and easily tunable to the reference data. Yet, there could be a temptation to switch to the models which are different from the drift-diffusion to enhance their validity range.

Where there is demand, there is supply. There are analytical models which were designed to capture ballistic transport in the nanoscale devices. In particular, the virtual source model [51] has gained a lot of momentum. This model operates with parameters which are essential for the ballistic transport (*e.g.* injection velocity) and has the extension towards the quasi-ballistic transport. Therefore, potentially it has a better predictive power than the drift-diffusion based BSIM-CMG. However, the virtual source model has a little number of additional features, like gate-induced drain leakage (GIDL) or self-heating effects, lacks fitting flexibility and it has not been widely adopted by the industry. Therefore, in this work MOSFETs are described using BSIM-CMG model extending it with quasi-ballistic transport as described in the next section.

2.2.2 Extending BSIM-CMG for the quasi-ballistic transport

Ballistic limit

Although it is true that for short channel devices velocity overshoot is a typical phenomenon, it does not lead to proportionally higher currents. This may be explained by the device performance at the ballistic limit. Figure 2.5 illustrates a typical band diagram of a MOSFET in the saturation regime.

Based on the simplistic gradual channel approximation [52], the inversion charge sheet density (Q_i) as a function of gate bias (V_{GS}) is expressed as

$$Q_i = C_{ox}(V_{GS} - V_{TH}), \tag{2.2}$$

where C_{ox} is the oxide capacitance, V_{TH} is the threshold voltage. Thus current on the source side would be just

$$I_{DS} = W_{EFF}C_{ox}(V_{GS} - V_{TH})v_s, \tag{2.3}$$

where W_{EFF} is the effective device width and v_s is the average carrier velocity near the source. To maintain current continuity, current on the drain side should

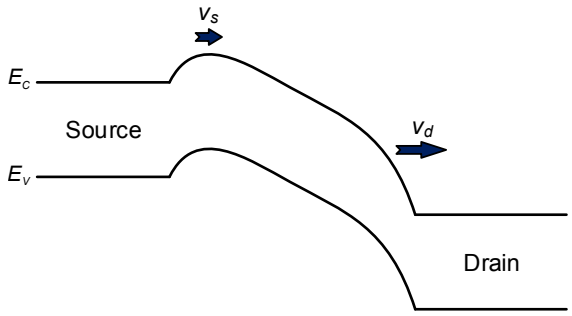


Figure 2.5: Band diagram of a MOSFET in the saturation regime; v_s and v_d are the average carrier velocities near the source and the drain, respectively. Adapted from [50].

be the same despite the velocity overshoot at the drain side. This balance holds true as the inversion charge sheet density is smaller at the drain side. That said, the performance limit comes from the *source* velocity. Its value is defined by the thermal injection velocity (v_{inj}) discussed in the ballistic MOSFET model by Natori [53]. Thus, for the fully ballistic transport, the saturation velocity term in the drift-diffusion model may be replaced by injection velocity. However, the transport in the advanced devices is quasi-ballistic, rather than fully ballistic. In addition, the injection velocity should be correctly computed or measured to make accurate current predictions.

Extension of drift-diffusion model for quasi-ballistic transport

If parasitic S/D resistances and short channel effects are ignored the fully ballistic current may be expressed with Eq. 2.3 if v_s is replaced with v_{inj}. We used the TCAD Synopsys Sentaurus Band Structure (s-band) top of the barrier simulations to extract both the number of inversion carriers $N_{inv} = WC_{ox}(V_{GS} - V_{TH})/q$ and v_{inj} as a function of applied voltage and stress in the channel. The latter is required to properly model transport in the lateral devices which typically have S/D stressors to boost current. The s-band program computes a subband structure of 2D devices for arbitrary surface orientation and strain by solving six-band $k\cdot p$ and two-band $k\cdot p$ Schrödinger equation for holes and electrons respectively [54]. In case of silicon, N_{inv} is weakly dependent on the applied stress and is defined by the gate voltage, while v_{inj} is not very sensitive to device bias, but is a strong function of stress (Fig. 2.6).

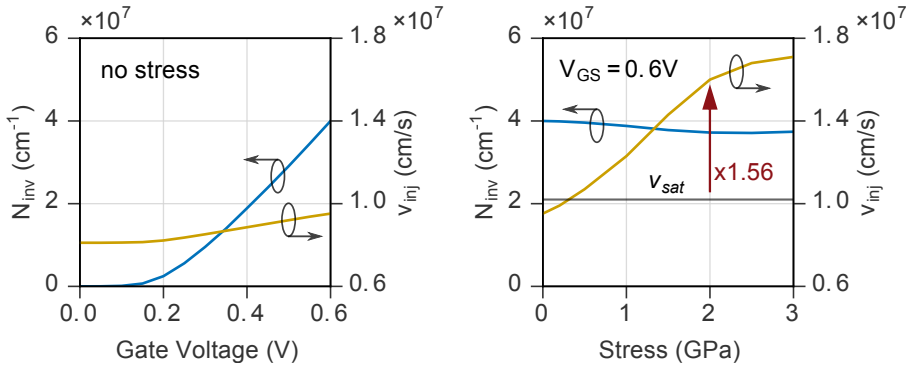


Figure 2.6: N_{inv} and v_{inj} as function of gate voltage (left) and applied stress (right) for a NMOS FinFET. Fin height is 30 nm and fin width is 5 nm. I_{OFF} = 3.5 nA, S/D resistances and short-channel effects are ignored.

Quantum confinement effects have an impact on the value of injection velocity and numbers of carriers in the channel and thus it has to be computed for various structures and stress levels. Yet, obtained values put an upper limit on device performance. In order to accurately compute the realistic drive values, it is necessary to assess the degree of ballisticity in the channel, or ballistic ratio (BR). It is a function of a gate length, and so is a device saturation current:

$$I_{DS}(L_G) = \mathrm{BR}(L_G) I_{\mathrm{bal}}, \tag{2.4}$$

where $\mathrm{BR}(L_G)$ is a ballistic ratio as a function of a gate length. Ref. [15] provides data on $\mathrm{BR}(L_G)$ for both stressed and relaxed FinFET devices. Difference in BR for various channel materials is discussed in [55]. With a FinFET scaling to a NWFET device surface to volume ratio changes, which may result in stronger scattering, which in turn lowers BR [56]. These literature data on BR are used afterwards as the basis for extending the BSIM-CMG drift-diffusion model into the quasi-ballistic regime. We will discuss this afterwards in chapter 4.

The question remains on how to capture a linear current correctly. A low field current is mainly defined by the carriers mobility. Thus the long channel mobility should be replaced by the apparent mobility due to ballistic mobility reduction and additional scattering mechanisms [44], [57], [58]. Apparent mobility may be calculated with Mathiessen's rule

$$\frac{1}{\mu_{app}} = \frac{1}{\mu_{long\ channel}} + \frac{\alpha_\mu}{L_G}, \tag{2.5}$$

where μ_{app} is an apparent mobility, α_μ is a mobility degradation factor and $\mu_{long\ channel}$ is a long channel mobility. Both a mobility degradation factor

and a long channel mobility depend on the device structure and the stress in the channel. However, it should be noted that α_μ cannot be lower than its minimum value given by the ballistic mobility — $\alpha_\mu = (2k_BT/q)v_{inj}$, v_{inj} being the injection velocity at the source as discussed earlier [44].

Modifications around BSIM-CMG

First, the compact model should be tuned to operate with the same charge as TCAD predicts. This may be done in BSIM-CMG by adjusting the position of the charge centroid. Proper V_{TH} is set by the work-function adjustment.

Second, the linear current modification are done through the apparent mobility introduction. We assumed the mobility reduction to be moderate [59] by setting $\alpha_\mu = 0.07\,\mathrm{nm\,V\,s/cm^2}$ for relaxed devices [60] and a bit lower $\alpha_\mu = 0.05\,\mathrm{nm\,V\,s/cm^2}$ for stressed devices [57]. This parameter is used in Eq. 2.5.

Last, the *VSAT* parameter has to be tuned to fit the actual saturation current from Eq. 2.4. As mobility gets lower with gate length scaling, saturation velocity is increasing similar to the work done in [61].

This methodology on the BSIM-CMG modification has been successfully tested. We created various compact models for different types of advanced devices and benchmarked them capturing the quasi-ballistic transport as described in the next chapters. All the modifications were made by redefining BSIM-CMG parameters with our own.

2.3 VFET parasitic resistances and capacitances

2.3.1 Modelling of vertical device parasitic resistances

Existing parasitic models for vertical devices

There are plenty of papers discussing analytical modelling of parasitic resistance in advanced devices, which may be used in compact models [62], [63], [64], [65], [66], [67], [68]. However, all the listed work dealt with the lateral symmetrical devices. Vertical devices are asymmetrical with a source resistance not being equal to a drain resistance. A contact to a bottom electrode needs to be done through a deep and narrow via, which translates into a highly resistive path (Fig. 2.7). As a result, this asymmetry has to be taken into account in parasitics compact model to enable accurate performance estimation.

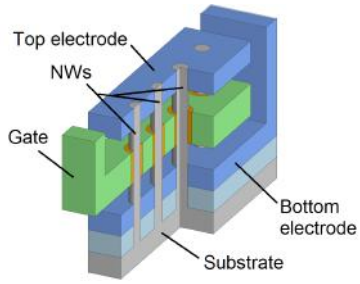


Figure 2.7: 3D sketch of a vertical device.

The work published in ref. [69] provides analytical models for the three major access resistance components of a vertical device: S/D extensions (ungated part of a device), the contact resistance of metal-semiconductor interface and electrode resistance. However, resistances are still treated in this work in a symmetrical way. An attempt to treat bottom and top electrode resistance differently is presented in ref. [31]. However, this work does not have a proper metallic bottom electrode. It relies on a doped substrate instead, which is likely going to be too resistive. On top of that the contact interface area between the metal plug and the substrate is going to be quite small. This again leads to a high access resistance. In this section we are discussing analytical modelling of a bottom electrode to semiconductor contact resistance accounting for metal resistivity, which is typically neglected in the conventional methods based on the transmission line model (TLM) [70]. The rest of the resistance calculations are based on the above mentioned references.

Bottom electrode connection

As we mentioned in the introduction, the focus of this dissertation is on the 5 nm node. This means that ground rules are heavily scaled leading to certain issues, which have been less pronounced so far. One of them is not anymore negligible metal resistance [9]. Let us explain the way we account for it.

There are different ways to draw layouts of standard cells based on vertical device but all of them are very much different from lateral layouts as all the layers are drawn on top of each other. As we will show in the next chapter, the layout shown in Fig. 2.8 is the one which has the best area efficiency, so we use it as the baseline for the parasitics model development.

Note the position of vias required for the bottom electrode connection. They are located either on the north or on the south. Current is injected through

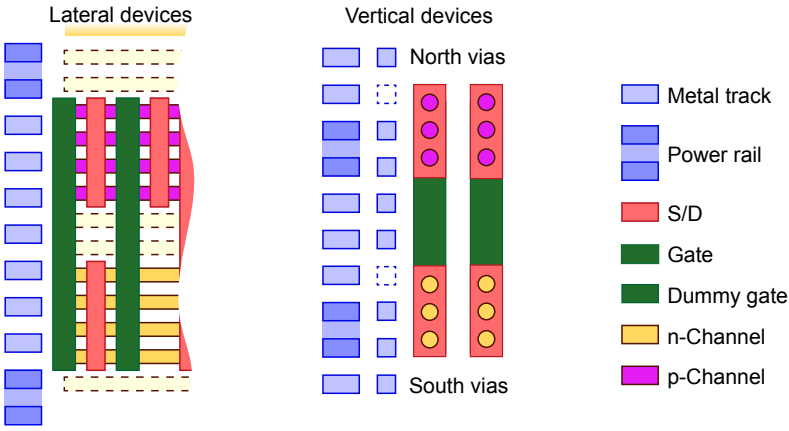


Figure 2.8: Layouts of lateral (left) and vertical (right) devices. Input is in the middle and output is either on the north or on the south [71].

these vias which are deep and narrow, and thus highly resistive. Next, current flows through the narrow metal around the channel (either a bunch of NWs as in the Fig. 2.8, or a single sheet). Due to the device pitch constraints, this metal has to be very narrow, so its resistance is inevitably high. It is hardly possible to fill the vias and electrodes with copper, so we assume that tungsten is used. Its resistivity properties are calibrated to the measurement data. Fig. 2.9a illustrates how tungsten wire resistivity changes with scaling based on our model which is just a mathematical fitting of the measurements. Fig. 2.9b demonstrates an impact on wire resistance. There is a about 3× increase in the resistance of a narrow wire with respect to the bulk resistivity case, so these increase should be taken into account in a model.

Model for contact resistance

In this section we present an analytical model of contact resistance between metal and semiconductor. This model is based on the transmission line model (TLM) method and was derived in analogous with ref. [70]. The model accounts for *metal resistance*, metal-semiconductor interface resistance and semiconductor resistance, and handles all the spreading components correctly. The validity of the model is verified through the comparison with the finite element modelling computations.

In the case metal resistance is neglected, an electrical potential is uniform across all the metal piece (Fig. 2.10). In this case the contact resistance may be

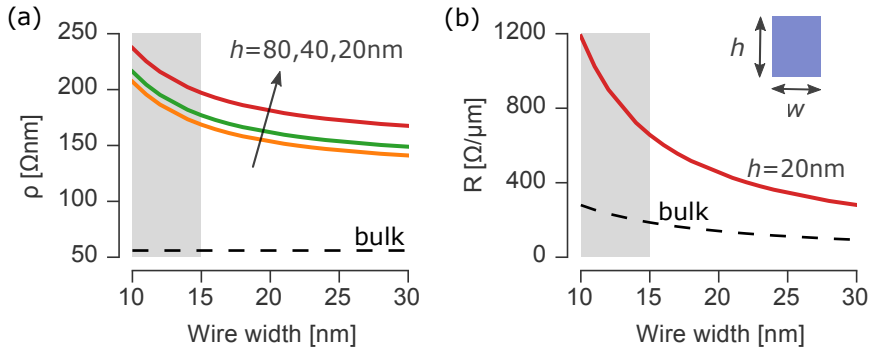


Figure 2.9: Tungsten wire (a) resistivity and (b) resistance as a function of wire dimensions. Increase in resistance for highly scaled wires is dramatic. Shaded region corresponds to the wire widths which will likely be used at 5nm technology.

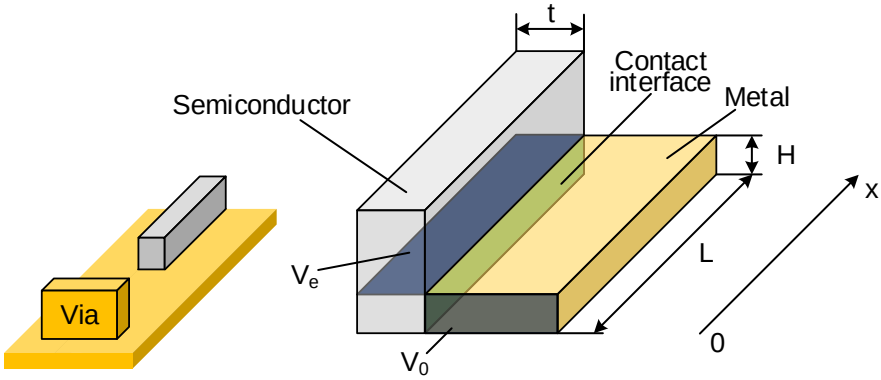


Figure 2.10: Left: the bottom electrode wrapping the elongated channel around has a via to the top placed by the short side of the channel. Right: the detailed schematics of the bottom electrode to the channel contact.

computed through the equation

$$R_{csimple} = \sqrt{\rho_c R_s}\coth\left(H/L_T\right)/L, \tag{2.6}$$

where $R_{csimple}$ is a simplified contact resistance (not accounting for a metal resistance), ρ_c is a specific contact resistivity, R_s is a semiconductor sheet resistance, H is a height of a contact, L is a length of a contact [72]. R_s may be computed as ρ_{sc}/t, where ρ_{sc} is a semiconductor resistivity and t – its thickness.

This formula accounts not only for the resistance of the interface, but also for the spreading resistance in the semiconductor with an assumption that potential V_e is the same along the blue plane shown in Fig. 2.10.

However, in case a metal resistivity plays a role, it is important to define where current is injected. In Fig. 2.10 it is indicated with the plane having the V_0 potential. We chose this plane based on the location of the bottom electrode access via (see layout in Fig. 2.8 and Fig. 2.10).

Now instead of analyzing the whole contact length, we focus on the infinitely small portion of it, dx. In this case Eq. 2.6 is still valid and may be written as

$$R_{cdx} = \sqrt{\rho_c R_s}\coth\left(H/L_T\right)/dx = B/dx, \tag{2.7}$$

where B is introduced for the simplification of the further calculations. Notice, that we make an assumption of constant potential in the metal cross-section at a given position x. This is true only if the semiconductor to metal interface is more resistive than metal resistivity, which is typically the case.

Now we can use the TLM method again but instead of ρ_c we use R_{cdx} from Eq. 2.7. We may write the following set of equations describing current flow and voltage distribution:

$$\frac{dI}{dx} = \frac{1}{B}(V(x) - V_e) \tag{2.8}$$

and

$$\frac{dV}{dx} = \rho_m I(x)/A, \tag{2.9}$$

where ρ_m is a metal resistivity, and A is a metal cross-section area. Following the ideas from ref. [70], we may obtain from Eq. 2.8 and Eq. 2.9 that

$$\frac{d^2I}{dx^2} - \frac{\rho_m}{AB}I = 0. \tag{2.10}$$

This equation is very similar to the results from the standard TLM, thus we may directly write an equation for the transfer length:

$$L_{Tm} = \sqrt{\frac{AB}{\rho_m}}, \tag{2.11}$$

where L_{Tm} is a transfer length accounting for metal resistivity. Next observation which comes from Fig. 2.10: current at the injection plate ($x = 0$) is i_0, however at $x = L$ current is zero because all the current flows into a semiconductor. With this assumption, we can write

$$I(x) = i_0\frac{\sinh\left(\frac{L-x}{L_{Tm}}\right)}{\sinh\left(\frac{L}{L_{Tm}}\right)}. \tag{2.12}$$

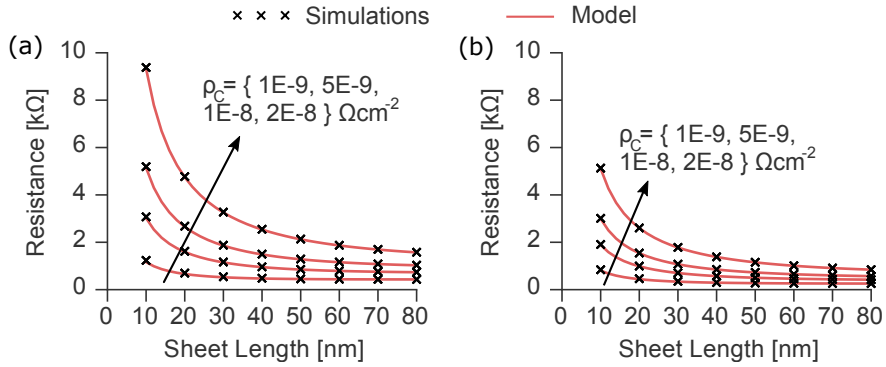


Figure 2.11: Match of contact resistance calculations between analytical solution and finite element modelling is very good for various specific contact resistivity values and contact lengths. In here, $H = 24\,\text{nm}$, metal width is 5 nm with resistivity of $2\,\mu\Omega\,\text{m}$, semiconductor has thickness $t = 5\,\text{nm}$ and resistivity $6\,\mu\Omega\,\text{m}$.

Finally we differentiate Eq. 2.12, combine it with Eq. 2.8 and obtain the expression for the contact resistance from plane V_0 to V_e accounting for resistivities of semiconductor, metal and interface together with all the spreading components:

$$R_c = \frac{B}{L_{Tm}} \coth\left(L/L_{Tm}\right) = \sqrt{\frac{B\rho_m}{A}} \coth\left(L/L_{Tm}\right). \tag{2.13}$$

The agreement of this analytical solution with the finite element modelling predictions is nearly perfect (Fig. 2.11a). It is relatively easy to adjust the calculations for the case of the double sided contact which happens in the reality: bottom electrode metal wraps semiconductor body from both sides (Fig. 2.10). In this case we have to divide R_c value by two (as we have a parallel connection of the resistors):

$$R_{c_{2sides}} = R_c/2. \tag{2.14}$$

On top of that, B and L_T should be computed for a half of semiconductor thickness. That said t should be replaced with $t/2$. Again, good match with the simulations is obtained as shown in Fig. 2.11b.

The next case to consider also comes from the actual layouts (Fig. 2.8). It is when current at $x = L$ is not zero but i_p, and, at $x = 0$, current is $i_0 = i_n + i_p$, where i_n is a current from nMOS device and i_p is a current from pMOS device (Fig. 2.12). For a good design, currents from both n- and pMOS are balanced,

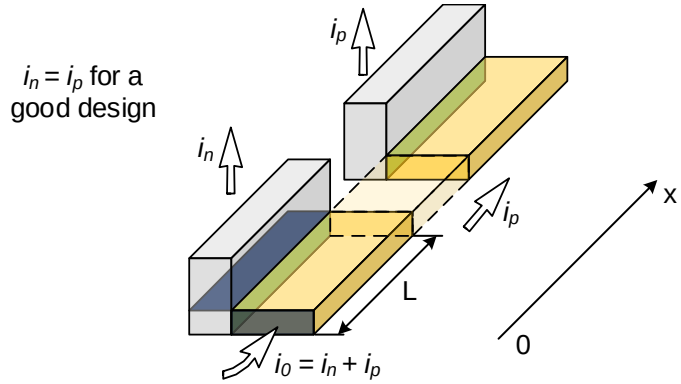


Figure 2.12: Schematics of the bottom electrode to the channel contact in case there are two devices in series sharing the same electrode.

that is $i_n = i_p = i_{np}$. For simplicity, let us introduce $a = 1/L_{Tm}$. General solution of the differential equation 2.10 is

$$I(x) = c_1 \exp(ax) + c_2 \exp(-ax), \tag{2.15}$$

where c_1 and c_2 are constants defined through boundary conditions, which are $I(0) = 2i_{np}$ and $I(L) = i_{np}$. After doing a bit of math, we may obtain these values:

$$c_1 = 2i_{np} - (i_{np}/2)[2\exp(aL) - 1]/\sinh(aL) \tag{2.16}$$

$$c_2 = (i_{np}/2)[2\exp(aL) - 1]/\sinh(aL) \tag{2.17}$$

Again, we differentiate Eq. 2.15, combine it with Eq. 2.8 and obtain the expression for the contact resistance from plane V_0 to V_e:

$$R_{c_{np}} = B(c_1 a - c_2 a)/i_{np} \tag{2.18}$$

We have also compared the results of the analytical modelling and the simulations. The matching is very good (Fig. 2.13). All of these analytical equations have been implemented in Verilog-A enabling accurate parasitic resistance assessment with the compact models.

2.3.2 Modelling of vertical device parasitic capacitances

Existing parasitic models for vertical devices

Capacitance affects not only speed of an IC, but also its switching energy. That is why it is especially important to accurately estimate parasitic capacitances

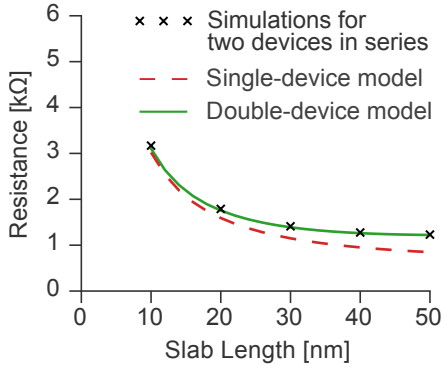


Figure 2.13: Match of contact resistance calculations between analytical solution and finite element modelling is very good for various contact lengths. We assumed metal thickness to be $H = 24\,\text{nm}$, metal width around the channel — 5 nm, metal resistivity — $2\,\mu\Omega\,\text{m}$, semiconductor thickness — $t = 5\,\text{nm}$ with resistivity of $6\,\mu\Omega\,\text{m}$, specific contact resistivity — $5{\times}10^{-9}\,\Omega\,\text{cm}^2$

for a given device. Just as in the case of parasitic resistance, most of the research has been focused on the lateral devices [64], [73]. We can pick up some interesting solutions from those models especially if they are formulated for the gate-all-around structures like ref. [74] or [75]. However, as we indicate in Fig. 2.14 that the available analytical model for circular NWs [75] does not work well for scaled diameters. The formulation of the model consists of two parts: relatively thin spacer region and relatively thick region. The transition between the two regions is not smooth as it should be. Moreover, in case we keep on using formulas for the relative thin spacers, they predict wrong results for thick spacers: capacitance increases with spacer thickness increase as shown with the dashed line. Therefore, we had to base ourselves on somewhat better models. Another reason not to adapt these models is that we did not want to limit ourselves to the nanowire-like shape of the channel, as we would also like to look at the slab-like channels.

There are some papers discussing parasitic capacitances in vertical devices specifically. For example, ref. [31] discuss parasitics specific to the channels made of vertical circular nanowires. Yet, it only assumes a single NW per device which is not practically useful due to the lack of drive current as we will show later. Another work explicitly mentions that it only provides a simple qualitative model [76], which is not enough for our study. Therefore, in this section we describe a generic model for the vertical devices capable of predicting parasitic capacitances for nanowire-like or slab-like channels for various number

of channels per device. On top that, this model accounts for possible impact of vias, either from the top to the bottom electrode or from the top to the gate electrode.

First, we observed that the difference in parasitic capacitance of NW-based VFETs marginally depends on the shape of NW channels: circular or square (Fig. 2.14). This small difference comes from the small NW size, which might be practically used: there is no room for NWs thicker than 10 nm in a VFET made with the design rules applicable to the 5 nm node. In this work, we continue with square NWs, as in this case the modeling approach becomes exactly the same for both NWs- and slab-like channels.

Elliptic integrals for parasitic capacitances

The overview of various parasitic capacitances in a vertical device is given in Fig. 2.15. Some of the capacitances, like an inner fringe capacitance are modeled by the core compact model, *e.g.* BSIM-CMG. A lot of other parasitic capacitances originate from the vias and therefore they are defined by an actual cell layout. All of these capacitances may be expressed through a composition of basic capacitances listed in Table 2.1. The parallel plate capacitance is easy to model and it is given with a final formula, but the other capacitances are modeled based on the methodology from [77], which relies on the calculation of the elliptic integral modulus k. Link between the modulus k and capacitance is

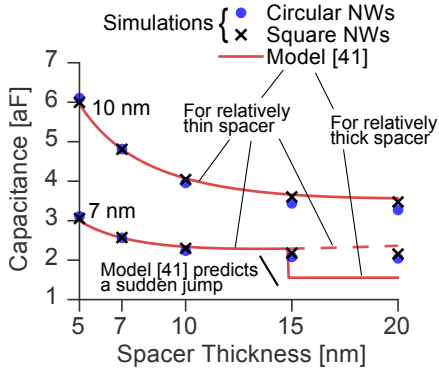


Figure 2.14: Parallel plate parasitic capacitance from the gate electrode to the S/D electrode **and** fringing capacitance to the S/D extension. The smaller the NW size is, the smaller the impact of its shape on parasitic capacitance.

given by Eq. 2.19 [78].

$$C = \begin{cases} \epsilon\frac{2}{\pi} \cdot \ln(2\sqrt{\frac{1+k}{1-k}}), & \text{if } k > \frac{1}{\sqrt{2}} \\ \epsilon\frac{\pi}{2} / \ln(2\sqrt{\frac{1+k'}{1-k'}}), & \text{if } k < \frac{1}{\sqrt{2}} \end{cases} \tag{2.19}$$

where ϵ is the dielectric permittivity and k' is the complementary modulus of the elliptic integral, defined as $k' = \sqrt{1-k^2}$. In order to compute k for various device geometries, these geometries should be conformally mapped to the $\Re$ axis of the complex plane. Once the mapping is done, the k values are obtained nearly automatically as explained in [77]. We perform the mapping with the holomorphic function $cos(y\pi/t)$, where y is the complex coordinate in the original plane and t defines the the width of the stripe in the original plane where capacitance should be computed. Essentially by using this function we decompose the VFET device into several strips, where capacitances are computed [79]. Table 2.1 provides a summary of the expressions for computing the elliptic modulus k for various basic capacitances. Notice that the k_{side} may be computed in an easier way as shown in Fig. 2.16. However, in this case, the electric field lines outside of the stripe region would also be included in the final capacitance solution which should be avoided with our partitioning scheme. Similar reasoning is applied for the case of k_{elec}. The next section is devoted to the description of the details regarding device partitioning so that the basic capacitances from Table 2.1 may be used.

Device partitioning

We partition the gate electrode into the small pieces as shown in the top-down view in Fig. 2.17. The green color indicates parallel plate capacitances, and

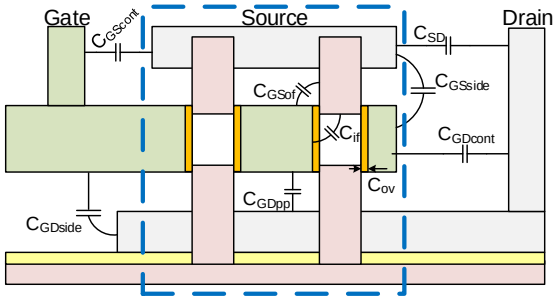
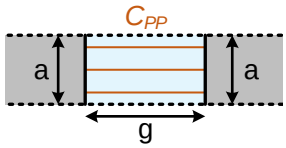
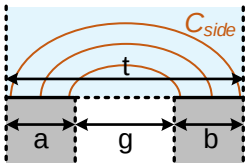


Figure 2.15: VFET cross-sectional view with the key parasitic capacitances. A lot of capacitances originate from the interconnects around the device.

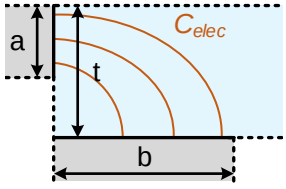
Table 2.1: Analytical formulas for the basic parasitic capacitances.



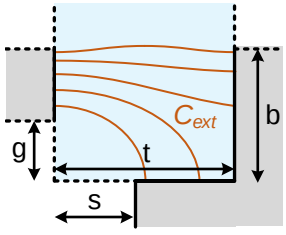
$$C_{PP} = \epsilon a/t$$



$$k_{side} = \sqrt{\frac{(1-\cos(\pi a/t))(\cos(\pi(a+g)/t)+1)}{(\cos(\pi a/t)+1)(1-\cos(\pi(a+g)/t))}}$$



$$k_{elec} = \sqrt{\frac{(1-\cos(\pi a/t))(\cosh(\pi b/t)-1)}{2(\cos(\pi a/t)+\cosh(\pi b/t))}}$$



$$k_{ext} = \sqrt{\frac{\cos(\pi s/t)+\cosh(\pi b/t)}{\cosh(\pi g/t)+\cosh(\pi b/t)}} \cdot \sqrt{\frac{\cosh(\pi b/t)-\cosh(\pi g/t)}{\cosh(\pi b/t)-\cos(\pi s/t)}}$$

Field lines are also computed in the shaded regions

$$k_{side} = \sqrt{\left(\frac{a}{g+a}\right)\cdot\left(\frac{b}{g+b}\right)}$$

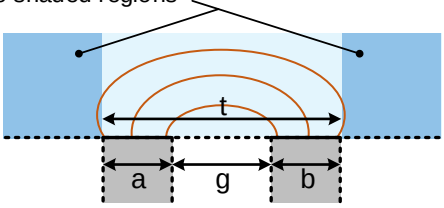


Figure 2.16: Another way of coputing k modulus for the side capacitance C_{side}.

the pink color indicates regions where both a parallel plate capacitance and a fringe capacitance to a S/D extension are calculated simultaneously. Light tone colors show that an extra attention should be payed to the edges of a devices.

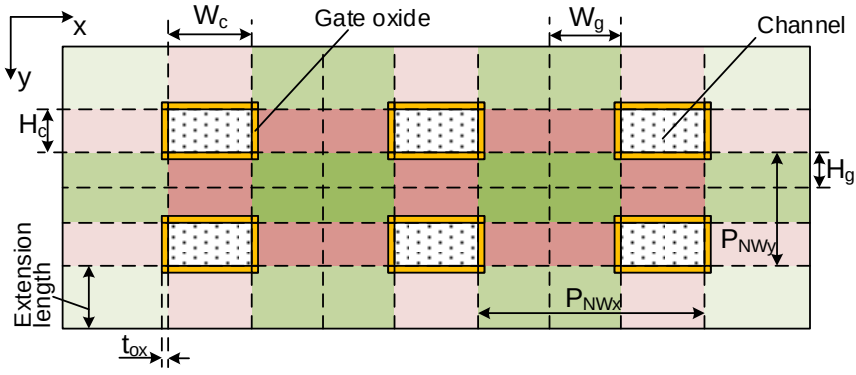


Figure 2.17: Top-down view (gate electrode section). The green color indicates the parallel plate capacitances, and the pink color indicates the region where both parallel plate and fringe capacitance to source and drain extensions should be calculated. Light tone on the sides shows that an extra capacitances between the sides of gate and source/drain electrodes should be included.

Some extra capacitances may originate from the fact that the electrodes may extend further than shown in Fig. 2.17. These electrode extensions are defined through the standard cell layouts.

Parameters like a number of NWs in x or y direction as well as a pitch (P_{NW}) between them are extracted from layouts relatively easy. However, the

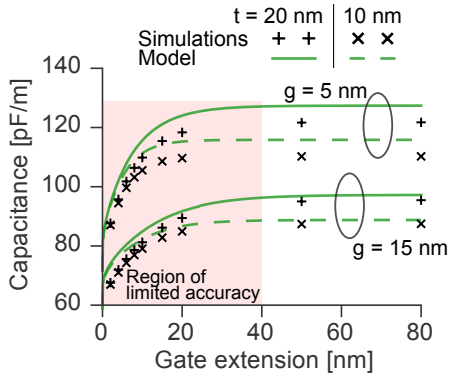


Figure 2.18: Capacitance as a function of gate extension beyond the edge of top electrode (see scenario "1" in Fig. 2.19) for various geometrical parameters. "g" defines the spacer thickness and "t" defines the thickness of the electrodes.

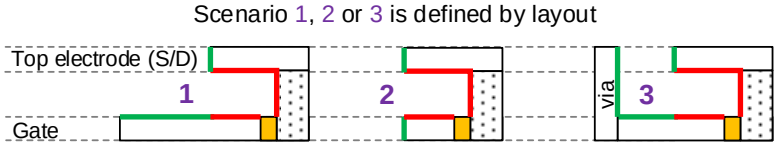


Figure 2.19: Different scenarios for the electrodes boundaries: one electrode may extend much further than than another one (1), they may have same extension length (2), one electrode may have a via to the top in the vicinity of another electrode (3).

extraction of electrode extension lengths and vias positions poses a challenge. A straightforward way is to just measure an extension from the NW edge until the electrode edge like it is defined in Fig. 2.17. This solution works most of the times, but not in the case when two devices share the same electrode (*e.g.* gate). In this case the extracted extensions are very long and their values are misleading. This scenario is labeled "1" in Figure 2.19. The solution is to clamp the maximum extension to the certain value beyond which the parasitic capacitance saturates and changes very weakly (Fig. 2.18). This defines the region where parasitics (not only capacitance, but also resistance) should be handled by the compact model but not with the PEX flow. Scenario "2" corresponds to the case of equal extensions for gate and top electrodes. It is tempting to use formulas for C_{side} to compute an extra fringing capacitance. However, there is an important detail in case the bottom electrode has the same extension length as both top and gate electrodes, the gate length should be divided by two in the capacitances calculations to account for the two side capacitances. In case there is a bottom to top via nearby, than the scenario "3" of Fig. 2.19 should be considered.

Models of capacitance components

Capacitances related to scenario "1" are modeled as a combination of C_{ext} and C_{elec} with a slight overestimation with respect to the simulation results (Fig. 2.18), which originates from the fact that individual capacitances are affected by the fringing fields from the neighboring capacitances. In addition, capacitance in the red region of Fig. 2.18 is not reliable, as an extra fringe capacitance to the side of the gate should be accounted for (see scenario "2"). However, in practice, the device configuration never corresponds to the dimensions from the red zone because the electrode extensions without the vias are only used to connect two separate devices which are on a large distance from each other (at least two metal pitches $= 2 \times 24\,\text{nm}$ for the 5 nm node).

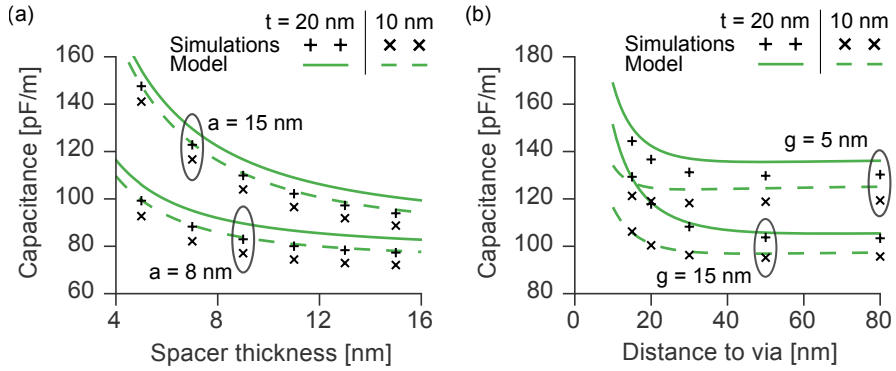


Figure 2.20: (a) Capacitances corresponding to scenario "2" for various geometries. "t" defines the thickness of the electrodes and "a" defines the size of the electrodes (extension from the edge of the NW). (b) Capacitances corresponding to scenario "3" for various geometries. "g" defines the spacer thickness and "t" defines the thickness of the electrodes.

The second scenario is widely used. The overall capacitance is a combination of C_{ext} and C_{side}. Our model predictions are quite well aligned to the simulations (Fig. 2.20a). Last scenario may be described with C_{ext} only, but used twice. The agreement with the simulations is satisfactory (Fig. 2.20b). Again, for both scenario number "2" and "3", there is a slight overestimation of capacitances, just as in the first case.

For all the scenarios, the inner C_{ext} capacitances may be replaced with C_{pp} if capacitance related to the light green region from Fig. 2.17 are concerned. An extra care should be payed to the corners in order not to compute parallel plate capacitance twice over there.

With all the basic components being described, we model parasitic capacitances of the full device in the three dimensional (3D) space and compare the results with the 3D simulations for various dimensions (Fig. 2.21). The slight overestimation related to the device partitioning is compensated by the corner effects. The accuracy of the models is very good across the wide range of dimensions without any fitting parameters used.

Case of drain underlap

The special case which mainly occurs in TFET devices is related to the drain underlap. We will discuss the advantages which the drain underlap structure

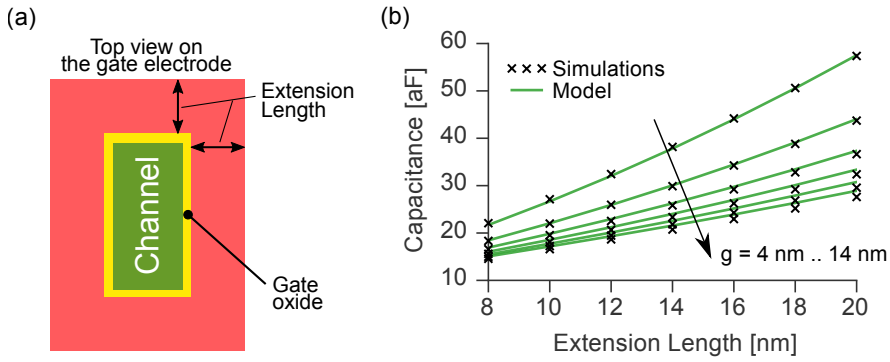


Figure 2.21: (a) Definition of extension length. (b) Comparison of parasitic capacitances from simulations and from analytical model for various geometries. "g" is the parameter which defines the spacer thickness.

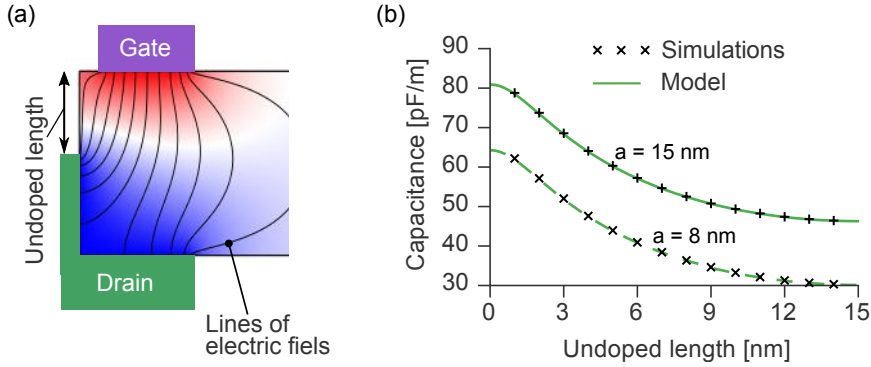


Figure 2.22: (a) Illustration of the drain underlap case: we do not compute capacitance from the gate to the channel as the latter is not doped. (b) Gate to extension and bottom electrode capacitance (as defined in (a)) dependence on the drain underlap length. 'a" parameter defines the size of the electrodes (extension from the edge of the NW)

brings in the chapter 5, but for now let us focus on its impact on parasitic capacitances. As part of the channel at the drain side is not covered by the gate as shown in Fig. 2.22a, the parasitic gate to drain capacitance is reduced. This happens because we do not compute capacitances from the gate electrode to the channel region as the latter is undoped. The reduction of the gate to drain capacitance is very important as it is this capacitance which is amplified

by the Miller effect. Therefore, it is of a high interest to model this capacitance precisely. In fact, all the formulas are already available: we just have to compute C_{ext} with an assumption that $s = L_{un}$, where L_{un} is the length of the undoped part. The comparison of simulations of a structure from Fig. 2.22a and our analytical model is given in Fig. 2.22b.

2.4 Summary and conclusions

This chapter discussed the importance of compact models in an IC design flow: an accuracy of the performance characterization largely depends on the quality of compact models used for the description of a device and related parasitics. Modelling of carriers transport in the channel of a device gets more and more sophisticated because of scaling. This means that models, like drift-diffusion, fail to accurately predict performance of the advanced devices. Yet, those models are well-established and commonly accepted by the industry. Therefore, in order to keep using them, we proposed a way to extend the drift-diffusion model to support the quasi-ballistic transport on the example of the BSIM-CMG compact model. However, it is not enough to accurately predict carriers transport in a channel to make conclusions on the device performance. Because of scaling, the role of parasitics (both resistance and capacitance) cannot be neglected in the advanced nodes. Moreover, these parasitics are dependent on the device structure. This works mostly deals with the vertical devices, therefore we proposed analytical models for both resistance and capacitance estimations. All the models were verified with finite-element modelling and coded in the Verilog-A language for further use in circuits simulations.

The key take-away messages from this chapter are the following.

- Drift-diffusion models may be tweaked to capture quasi-ballistic transport by modifying carriers mobility and saturation velocity. We proposed a rather simple method to do this on the example of the BSIM-CMG compact model.
- Device *RC*-parasitics may be described analytically. This approach allows to easily see the impact of different cell layout styles on the device performance, and ease device design optimization. We developed a parametrized *RC*-parasitics model for vertical devices.
- A model which accurately describes a carriers transport in the channel of a device along with a model of the corresponding device *RC*-parasitics enables efficient device optimization and characterization in SPICE, which is the focus of the next chapters.

CHAPTER 3

Vertical Layouts

3.1 Overview of lateral layouts scaling

The lateral layouts should be more or less familiar to the reader, yet we believe it is essential to make a few comments on them. As we briefly explained in the introduction, the cost of every new technological node dramatically increases, which means that the conventional 0.5× scaling of gate pitch and metal pitch product might be not sufficient to justify the new technology adoption from the economics point of view. No matter how good the technology is, it should provide a positive return on investment (ROI). There are two ways out. First, more than 0.5× scaling. We can scale the ground rules down to the level when we achieve conventional 0.5× area scaling and then apply extra tricks to scale area even further [80]. However, it might be technically easier (and, thus, more economically reasonable) to push ground rules just a little bit, and achieve the conventional 0.5× scaling with the same extra tricks as in the first scenario. This is the nowadays trend [8]. Probably, the most important trick here is a cell height reduction (Fig. 1.1). For lateral devices, this concept is linked with the aggressive fin pitch scaling or/and fin depopulation [20], [80], [81], [82]. The main idea is to drop a fin to gain area. The performance is maintained by making the remaining fins taller.

We start our analysis by looking at the 7 nm technological node to establish the performance reference point. At this moment of time, it is clear that this node will be based on lateral devices. Arguably, these lateral devices will be FinFETs or NWFETs. As a reference for the 7 nm node we pick the 7.5 tracks tall cells [8], [83]. Following the Intel approach [80], the likely 5 nm node would consist

Table 3.1: Ground rules for lateral devices at 7 nm and 5 nm technological nodes.

Node	CGP [nm]	MP [nm]	FP [nm]	Cell height	Number of fins
7 nm	42	32	24	7.5	3
5 nm	32	24	24	7.5	2
				6.0	1

of 6.0 tracks tall cells made with scaled ground rule to achieve more than 0.5× scaling. Yet, we also look at just 0.5× scaling keeping same cell height (7.5 tracks). Summary of the assumed ground rules for these two nodes is given in the Table 3.1. These rules are more or less consistent with the industrial scaling path (Fig. 1.2). Notice that the fin pitch is not scaling going from 7 nm to 5 nm as otherwise it would hit the limit of the self-aligned quadrupole patterning (SAQP) in the 193 nm immersion lithography [84], and it would pose severe challenges on the gate stack fabrication [85].

It is important that we have fixed the layouts of lateral devices as we will use them to find the vertical device dimensions which would enable iso-area comparison of two architectures. We can already imagine that the competition at iso-area will be hard for the vertical devices as lateral devices are extremely small already. Nevertheless, let us have a look at various layout options for the vertical architecture.

3.2 Layouts

3.2.1 Pin accessibility *vs.* area efficiency

Layouts of VFETs-based standard cells are different from the layouts based on lateral devices. They are difficult to interpret, because from the top-down view all the electrodes (top, gate, and bottom) overlap. Due to this fact, a vertical *device looks* very compact. However, it is necessary to contact all of the device electrodes. The only way to do this is to place the vias outside of the device boundaries, which increases its footprint.

Not that much work has been done on the optimization of layouts based on vertical devices. Ref. [29], [86] propose the layouts which look quite similar to the layouts made with lateral devices. Early work of T. Huynh-Bao [32] proposes to use the inbound power rails to improve the S/D and gate accessibility. With

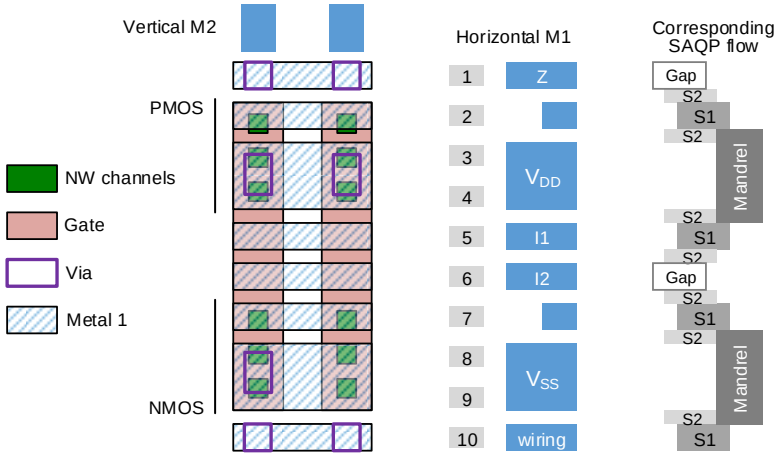


Figure 3.1: NAND2 layout with the inbound power rails. To simplify the picture, just a few layers are shown.

this idea in mind we tried to make some device geometry optimization in [87]. It turned out that it is still more efficient way to configure the channels in a device in a shape of a single column of NWs or, maybe, a nanosheet. This finding is correct for the digital logic, which is the focus of this work. In order to optimize the RF properties of vertical devices, quite different geometries should be chosen [69]. Probably, the most comprehensive study to date was published in [88] as it discusses different layout options and conducts the analysis in the scaled dimensions accounting for the lithography patterning constraints. We will base the following discussion mostly on these results.

Inbound power rails

In the case when a vertical device consist of several NWs placed in a column, it is quite logical to use the inbound power rails to free more routing resources. In this case the middle of the cell might be used for the intra-cell routing as well as for the input gate connections. The source and the drain electrodes are connected either from the top or from the bottom of the cell as shown in Fig. 3.1. The difficulty arises in the case if a series connection is needed between the transistors [86] as either the orientation of source and drain in transistors should be changed, or deep extra connections running from the top to the bottom are required.

One might want to place a connection to a bottom electrode or/and to a gate

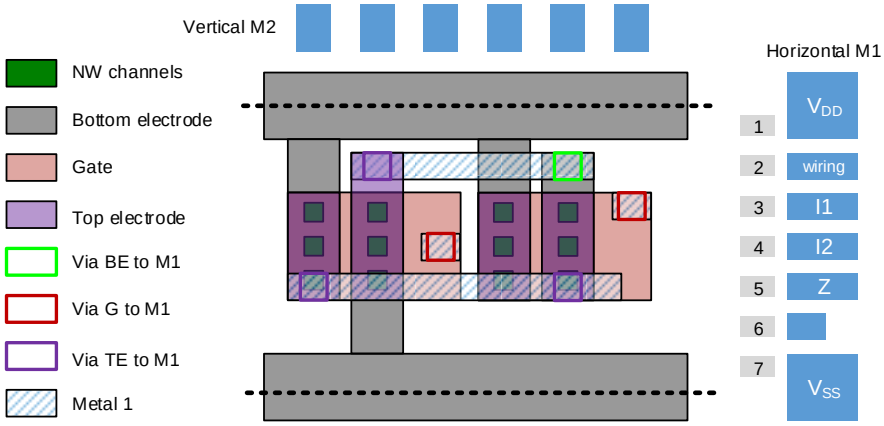


Figure 3.2: NAND2 layout with the interleaved diffusion. To simplify the picture, not all the layers are shown.

electrode on the left and on the right from the cell. Such a layout style has two drawbacks. First, the pitch between the adjacent devices largely increases. Second, the shape of the electrodes becomes quite complex, which makes it nearly impossible to print. In contrast, the layout template proposed in Fig. 3.1 is based on the 1D lines, which simplifies the patterning flow. The pitch between the devices is defined by the body thickness of the channel and the gate stack thickness. The main drawback of the proposed template is a poor pin accessibility. Only one vertical metal line (Metal 2 level) passes on top of the PMOS-NMOS device pair, which makes it difficult to propagate signals to the top metal layers.

Interleaved diffusion

In this section, we will discuss another layout template which improves the port accessibility. The previously discussed template still somehow resembled a conventional CMOS-based lateral layouts in a sense that all the PMOS devices were on the top and all the NMOS devices were on the bottom of the cell. For the lateral devices, this solution is quite common because of the necessity to share the diffusion. However, the vertical architecture does not require this. PMOS and NMOS device may be reshuffled to improve the layout density (Fig. 3.2). We tried to put these devices adjacent one to another one. The power rails in this case is outbound and it is shared between the adjacent cells. Again, there are three tracks for the intra-cell routing and gate connection and

two cells outside of device for the S/D connections. As the NMOS device now is on the left from the PMOS device, the height of the standard cell is reduced. The price to pay is an increased cell width. This comes with an advantage of better port accessibility because much more vertical metal lines intercept with a cell.

Just by looking at the proposed layouts of the NAND2 cell, we may see that the layout based on the template with the interleaved diffusion is about two times larger than the one based on the template with the inbound power rails. The exact area difference differs from cell to cell, but it is always quite significant. Therefore, we choose the template from Fig. 3.1 for the further analysis. Yet, we have to keep in mind that it might happen that at the place and route level of a big chip the difference in area between the two templates will be reduced because of the poor port accessibility of the chosen template.

3.2.2 Interchangeable source and drain

A conventional lateral device has a source equal to a drain (unless it is a tunnel FET device). However, a vertical device is intrinsically asymmetric. As all the connections in a chip are comping from the top, the top electrode is close to the interconnect network than the bottom electrode. It is the top to bottom via which makes the difference. Although this via is metallic, it is still highly resistive due to its high aspect ratio and narrow cross-section. The depth of the via is a sum of a gate length and of two spacer thicknesses. As both these parameters are not constrained by area, they might be quite relaxed. Thus the total via depth might easily reach 50 nm or be even longer.

At the same, the via cross-section is very small. In the case MP = 24 nm, its width is only about 12 nm. In the other direction it should be at most the device width, or about 25 nm as we will show later. These high aspect ratio vias are impossible to fabricate in copper, therefore we rely on tungsten. We will further discuss the impact of this resistive via later on in the thesis. Right now, we just want to stress that it is not negligible and it leads to the device asymmetry. Therefore, we have to decide whether there is a preferred device configuration.

In the case of the inbound power rails layout template, the V_{DD} and V_{SS} connections are in the vicinity of the device top electrode. The bottom electrode, however, is only accessible through the highly resistive via. On top of that, as the contacting to the bottom electrode is done from its short side (Fig. 3.1), there is a highly resistive path between the first and the last NW in the line. We have already discussed this issue in the previous chapter when we described the resistance model of the bottom electrode. Therefore, to capitalize V_{GS}

in VFETs, source is placed at the top electrode and drain is at the bottom electrode.

3.3 Area of vertical and lateral devices

There are different ways to compare the PPAC metrics of various technologies. In this work, we will pursue the one where the footprint area is fixed as this ultimately defines the cost. The two remaining variable to compare in this case are power and performance.

To ensure the comparison of vertical and lateral architectures under the same footprint, a vertical device pitch (VDP) (a vertical equivalent to CGP) should be linked with a CGP of lateral devices with Eq. 3.1.

$$\mathrm{VDP} = \frac{H_l}{H_v}\frac{N+1}{N}\mathrm{CGP}, \tag{3.1}$$

where H_l and H_v are the cells' heights expressed in number of tracks for lateral and vertical architectures, respectively; and N is a cell's width in a number of stacked devices. The extra unity in the numerator accounts for dummy gates in lateral devices which are used to separate fins between adjacent cells [83]. In case of NAND2 cell, $N = 2$. We believe that NAND2 is a good choice for area comparison as inverter is not representative (vertical inverters are much more compact than lateral inverters), while more complex cells, like D flip-flop, may be designed in a lot of different ways. Moreover, in [32], we have already demonstrated that VFETs are beneficial for complex designs (e.g. a VFETs-based 32-bit multiplier is 19% smaller compared to its lateral version), while still starting with an assumption of the equal lateral and vertical NAND2 cell area.

Table 3.2 indicates what should VDP be in order for VFETs to be competitive with lateral devices in area at 5 nm node. The number of tracks in vertical devices is directly linked with a number of NWs per device so the larger it is the better in case the overall performance is drive current limited. On top of that, tracks number cannot be chosen randomly because of restrictions coming from the lithography patterning rules.

Let us see whether there are any limitations on the VDP which would impact the choice of a particular vertical device cell height. As we will show in the chapter 4, diameter should be about 7 nm for the NW-like channels. The realistic gate dielectric for the advanced technological nodes like 5 nm is composed of 0.5 nm thick SiO_2 and 1.5 nm thick HfO_2. The gate p-metal thickness is around 2 nm

Table 3.2: VDP values (in nm) resulting in similar vertical and lateral cell footprint as a function of vertical and lateral cells height. Lateral CGP is 32 nm. Green color indicates the selected values.

Lateral cell height	Vertical cell height		
	12 tr.	10 tr.	8 tr.
7.5 tracks	30	**36**	45
6.0 tracks	24	29	**36**

and n-metal thickness is around 6 nm [85]. As the gate should surround the wires from all the sides, the total thickness of the gate electrode becomes 27 nm as seen from the top. Having VDP = 32 nm at 5 nm node means that spacing between the adjusting devices is only 5 nm, which is not good from the parasitic capacitance point view (gate to gate capacitance) not to mention difficulties in manufacturing. Therefore, it makes sense to increase VDP up to 36 nm as indicated with bold font in Table 3.2.

3.4 Summary and conclusions

A few VFET-based layout templates were reviewed in this chapter. In order to compete with lateral layouts in terms of area, extremely small vertical devices are needed, which poses a question whether they have sufficient drive current to compete with lateral devices in terms of performance. The stringent area requirement push us to choose a compact layout template despite its poor pin accessibility which might be an issue at the place and route level. We have also discussed a device asymmetry issue which is not present for lateral devices. In the case of the inbound power rails, it is better to place source on the top of the device to minimize the IR drop on the access resistance of the bottom electrode. Therefore, we will mostly use this terminals orientation in the further work.

The main conclusion of this chapter is that the footprint of vertical devices is not that much smaller than that of lateral devices because of the need to have connections to the gate and bottom levels: there should be escapes somewhere and they consume area. We proposed quite an area-efficient layouts with these escapes positioned on the north and the south of the devices. These layouts are EUV-ready but also compatible with the SAQP technique in the 193i lithography.

CHAPTER 4

Vertical MOSFET

4.1 Nanowire or nanosheet channel?

This section provides the comparison of MOSFETs with vertically oriented channels having NW-like shape or nanosheet (NSh)-like shape (Fig. 4.1, left). First, in this subsection, we share the general thinking process which led us to these channel options. Next subsection is devoted to the detailed comparison of these options on the DC level. Last subsection is about the benchmarking on the ring oscillator level.

The scaling path towards NW MOSFET has already been discussed in details in the introduction. In here, we provide just a short summary for the readers convenience. Every new technological node should provide improvements in the PPAC metrics to justify its adoption. With more and more mobile devices in out life, energy savings get more important than performance gains. Energy per switch is directly proportional to capacitance so it is of the utmost importance to lower capacitances, both on the device level and on the back end of line (BEOL) level. Yet, the focus of this thesis is only on the device optimization. Talking about device capacitances, we may distinguish between two capacitive contributors: channel capacitance and parasitic capacitances. Channel capacitance, as the name suggests, is defined by the channel structure and the equivalent oxide thickness (EOT) of the gate insulator. In order to compare channel capacitance of various Si MOSFETs it is sufficient to compare their effective width if EOT and gate length are fixed the same. It is the first order comparison, but still quite useful in practice.

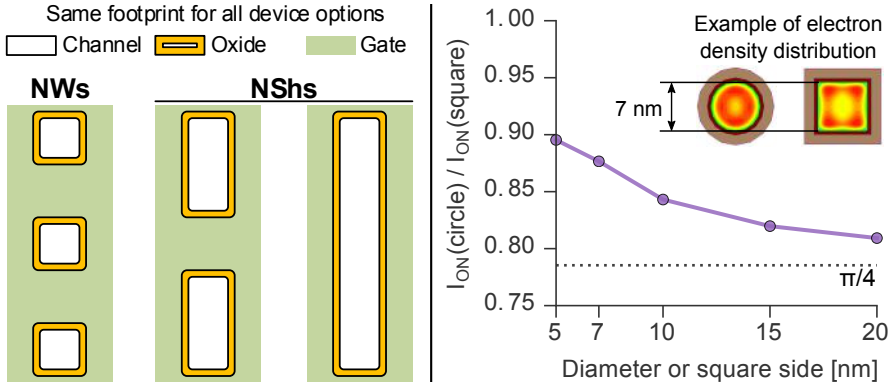


Figure 4.1: Left: Top-down view at the gate electrode level clearly demonstrates the different between NW-like channels and NSh-like channels. Right: NW shape has in impact on device performance, but its getting less important with scaling because of beneficial impact of confinement on electron density for circular channels (see example for the 7 nm thick channels, simulation results are from the Synopsys Sentaurus s-band simulations).

Next, current coming out of a device is also proportional to its effective width. As the switching delay is simply $\tau = CV/I$, one can get a perception that with device width scaling, delay remains constant but power consumption drops. That would be an ideal case. In reality, both device parasitic capacitances and BEOL interconnects do not necessarily scale with the device effective width. Thus, we risk to end up in a situation when device is not capable of driving these capacitances as fast as the non-scaled device.

This problem may be solved relatively easy in both FinFETs and lateral NWFETs as they have an ability to modulate their effective width (and, thus, drive) without changing their footprint by adjusting their fin height or number of stacked NWs, respectively. In order to change drive of a vertical device one has to change its footprint. This is a serious concern regarding vertical architecture as area savings is the must have condition for the technology adaption.

As a NW MOSFET allows to scale the gate length most aggressively thanks to its ultimate electrostatics control of the channel, we start with this option. A short gate length results in a small channel capacitance; yet due to a limited effective width of a NW MOSFET, it may fail to provide sufficient current. A possible way out applicable to the vertical architecture without area penalties is to replace several NWs with a NSh having a larger effective width. The degraded electrostatics may be compensated by a longer gate as increase in a gate length of a VFET may be done within the same device footprint. Talking

about effective width it is important to agree on the shape of various devices: they may have some roundings in the corners. This is not really important for a NSh channel as its effective width is mainly determined by the its length. However, a NW essentially consists of four corners, which brings us to two extremes: a circular shape or a square shape. A circular NW has $\frac{\pi}{4}\times$ effective width of a square NW. In case the body dimensions are large, the difference in full ballistic current approaches this difference in effective width (Fig. 4.1, right). However, with scaling this difference diminishes. This may be attributed to the fact that because of the higher confinement a circular NW has higher carriers density than a square NW as shown in the example for the 7 nm thick channels. Therefore, for the dimensions of interest (diameter less than 10 nm), the difference in ballistic current for circular and square NWs is utmost 15%. As it is hard to predict which exact shape a channel would have once fabricated, for our simulations we use square channels with 1.5 nm rounding of the corners. For NShs we start from rectangular channels but apply the same rounding to the corners.

To summarize, we *expect* vertical devices with NWs channels to consume least power but to be not suitable for driving heavy loads. In turn, nanosheet channels should provide a lot of current which may be especially interesting for applications like field-programmable gate array (FPGA) having long interconnects. The price to pay for a high drive — an increased energy consumption.

4.2 DC performance

4.2.1 Ballistic Current

The device dimensions may be extracted from the layouts of the VFET-based cells discussed in the chapter 3. For the reader's convenience, the reference simplified layout is repeated in this section (Fig. 4.2). Both nMOS and pMOS consist of just three NWs per device. How much current can we get out of these devices? In order to properly answer this question, let us build up the device model step by step starting with the understanding on the ultimate (ballistic) current limit which we may expect from a single NW or a NSh.

In case there is no access resistance, current predictions for vertical devices should not be any different than for lateral devices. Yet, it is not exactly true. Vertical NWs stick out of wafer perpendicular to its surface. The typical wafer orientation for lateral devices is (001) with a channel formed along the [110] direction. This means that in case the same wafers are used for a vertical device,

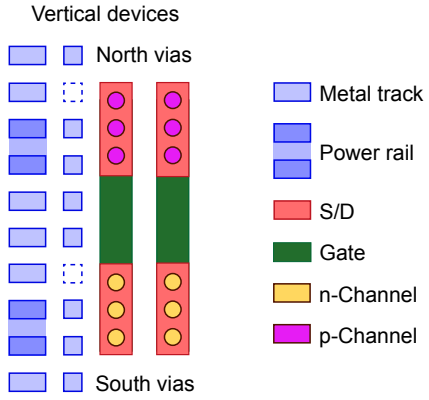


Figure 4.2: Simplified sketch of a ten tracks tall VFET layout.

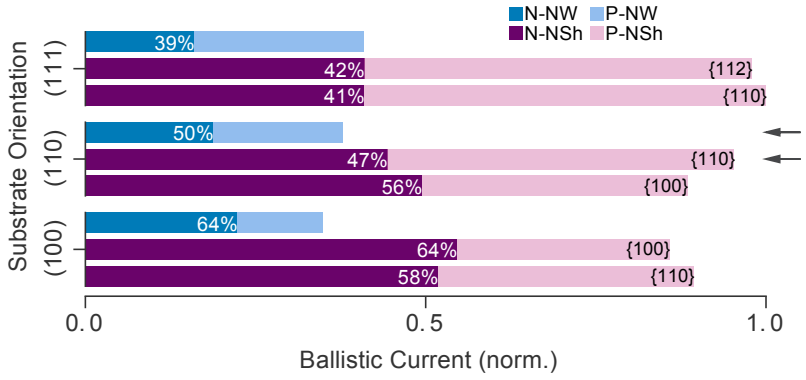


Figure 4.3: Ballistic nMOS and pMOS currents for NW and NSh for various substrate orientations. NW has a diameter of 7 nm; NSh size is $30 \times 5\,\mathrm{nm}^2$. Sidewall orientation for NSh is given in curly brackets. Percentages indicate a fraction of the nMOS current to the sum of nMOS and pMOS currents. All the devices have equal off-current.

its channel is [001]-oriented. Channel orientation difference causes difference between lateral and vertical device performance. How to capture it?

As we explained in the second chapter, we assess various devices taking into account the quasi-ballistic carriers transport in the channel. The first step towards this transport model would be an estimation of the full ballistic transport. We used the TCAD Synopsys Sentaurus Band Structure top of

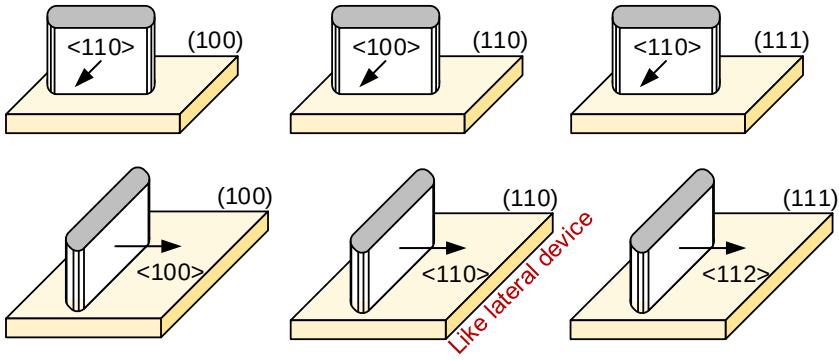


Figure 4.4: Six orientations of a nanosheet channel are considered through the combination of the various substrate orientations and the rotation of nanosheets around the transport direction.

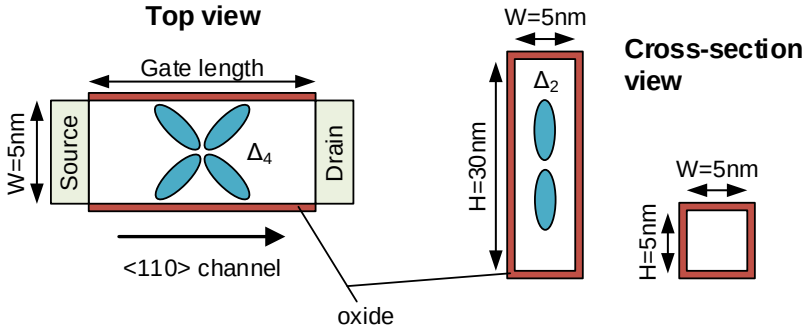


Figure 4.5: Respective orientation of Si nMOS Δ_2 and Δ_4 valleys to the device geometry.

the barrier ("s-band") simulations to compute it. This program solves the Schrödinger equation on 2D devices for *arbitrary surface orientation* accounting for all the quantum confinement effects [54]. S-band uses the parabolic Schrödinger equation with a correction for nonparabolicity for the silicon conduction band and the six-band $k \cdot p$ Schrödinger equation for the silicon valance band. This should be sufficiently accurate for our purposes.

Figure 4.3 provides a summary on ballistic currents for various substrate (and, thus, channel) orientations. The length of the bars corresponds to the sum of nMOS and pMOS ballistic currents, normalized to the highest value. Blue color

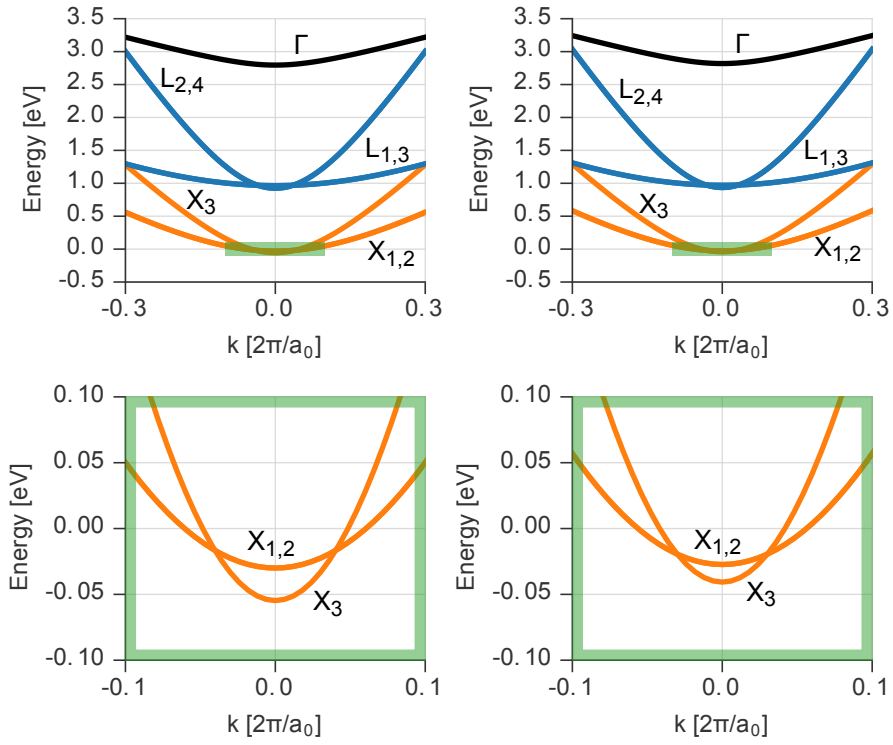


Figure 4.6: Band structure of 5x5 NW (left) and 30x5 NSh (right) nMOS devices. Only the lowest subband per band is shown. Bottom row is a zoom-in of X-valleys indicating larger band-bending for 5x5 NW.

is for NWs and purple is for NShs. NShs sticking out of the wafer may be rotated differently as shown in Fig. 4.4. The curly brackets indicate the orientation of their sidewalls. Percentages indicate a fraction of the nMOS current to the sum of nMOS and pMOS currents. The preferred orientation would be defined not only by this sum, but also by the balance between nMOS and pMOS drive as it might be quite tricky to balance them otherwise. Therefore, although the (111) substrate provides the best overall performance, we use (110) substrates with (110) NSh sidewall to have well balanced nMOS and pMOS devices. As for NWs, the rotation around the transport direction does not change anything, as devices are symmetrical. Again, in order to balance nMOS and pMOS devices, we choose (110) substrate orientation for the NWFETs.

Next, we should make a detailed comparison of the ballistic current from

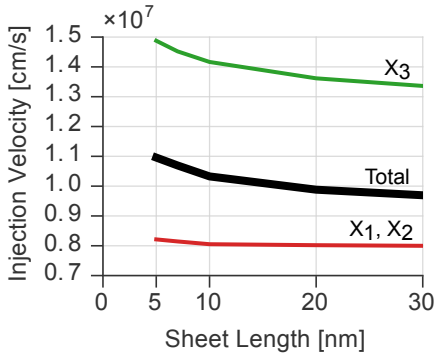


Figure 4.7: Electron injection velocity increases especially fast for the X_3 valley with sheet length scaling.

NWs and NShs. We have mentioned in the overview that current is roughly proportional to the device effective width. For the NSh case from Fig. 4.3, $W_{EFF} = 70\,\text{nm}$, while for the NW $W_{EFF} \approx 22\,\text{nm}$. Thus, we may expect a NSh to provide about 3× higher current. However, according to Fig. 4.3, a NSh is only 2× stronger than a NW. One may say that NW channel provides better electrostatics control than NSh. This is true, however the s-band tool computes ballistic current using top of the barrier approach, so it is fair to say that the assumed gate length is infinitely long, which results in the ideal electrostatics for both NW and NSh (subthreshold slope is about 60 mV/dec). Therefore, the true explanation is related to the difference in the subband structure induced by the quantum confinement effects.

For the nMOS devices, this may be explained with Figure 4.5. With sheet length scaling, quantum confinement effects increase the subband energy of the Δ_4 valleys more than the Δ_2 valley because the latter has heavier quantization mass in the height direction. This results in the increased occupation of the Δ_2 valley and consequent drive current increase. The reason behind is that the transport mass of the Δ_2 valley ($0.19m_0$) is much smaller than that of the Δ_4 valleys ($0.55m_0$), which in turn results in increased thermal injection velocity [56]. m_0 here is the electron effective mass.

Thus, we expect the injection velocity to increase with the transition from NSh to NW. Figure 4.6 indicates the lowest subband band structures for nMOS NW and NSh based on the s-band simulation results. The conduction is completely dominated by X bands. Moreover, as expected, due to the geometrical confinement, X_1, X_2 and X_3 valleys are not the same for NSh and NW, which results in different injection velocity per valley (Fig. 4.7). Another

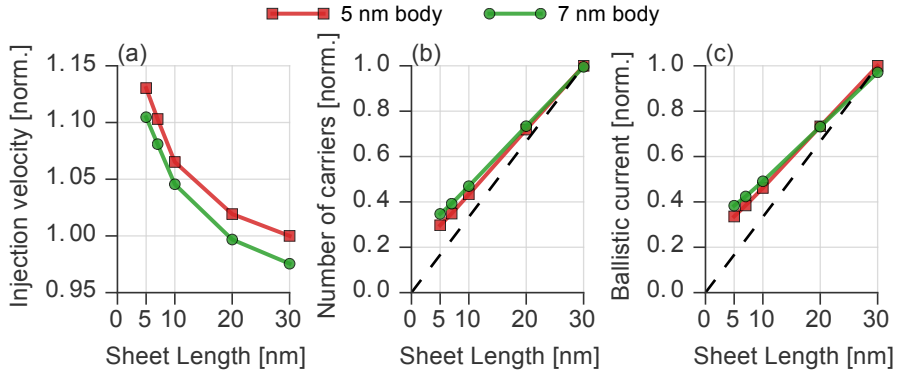


Figure 4.8: Impact of sheet scaling on a) electron injection velocity, b) number of inversion carriers, and c) ballistic current. All values are normalized to the longest sheet (30 nm long). For all the devices, the off-current was adjusted to the same value (10 nA) with the work-function tuning.

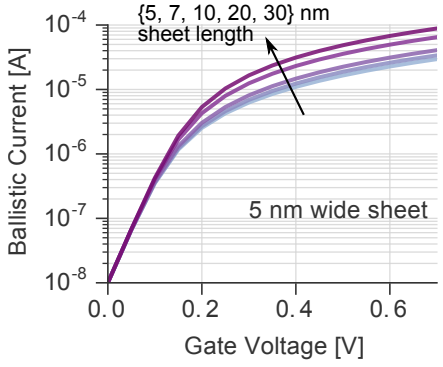


Figure 4.9: Ballistic $I_{DS} - V_{GS}$ characteristics for various devices. The work-function was adjusted to target the same off-current (10 nA).

difference related to the confinement is in the number of inversion carriers contributing to the current per unit channel width. Both increase in injection velocity and carriers density explains why the ballistic current per channel width in NWs is higher that it is in NShs. Figure 4.8 provides an overview on the transition from a NSh to a NW. Note, that the off-current was adjusted to the same value (10 nA) by the work-function tuning for all the devices to ensure the fair comparison (Fig. 4.9). This adjustment is also beneficial for a NW channel.

Similar explanation of higher currents for NWs may be given for the pMOS

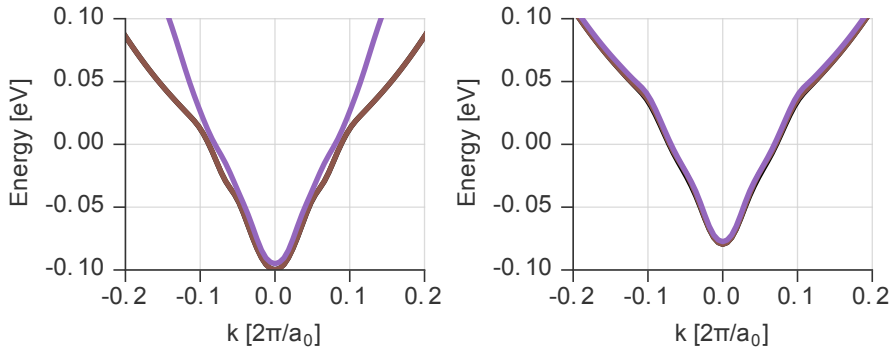


Figure 4.10: Three bottommost valence bands for 5x5 NW (left) and 30x5 NSh (right) pMOS.

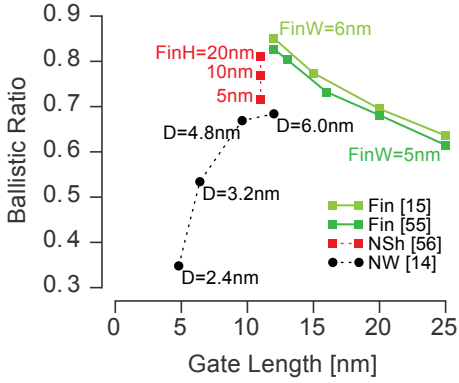


Figure 4.11: Summary of available literature data on ballistic ratio for Si-based multiple-gate devices.

devices, although the valence-band structure of Si is more complex than its conduction-band. Figure 4.10 shows the three bottom valence subbands out of 256 subbands considered in the simulations. Scaling of NSh leads to more curved bands, which in turn results in higher injection velocities for NW-like structures: $v_{inj} = 1.13{\times}10^7$ cm/s for the $5 \times 30\,\mathrm{nm}^2$ NSh *vs.* $v_{inj} = 1.20{\times}10^7$ cm/s for the $5 \times 5\,\mathrm{nm}^2$ NW.

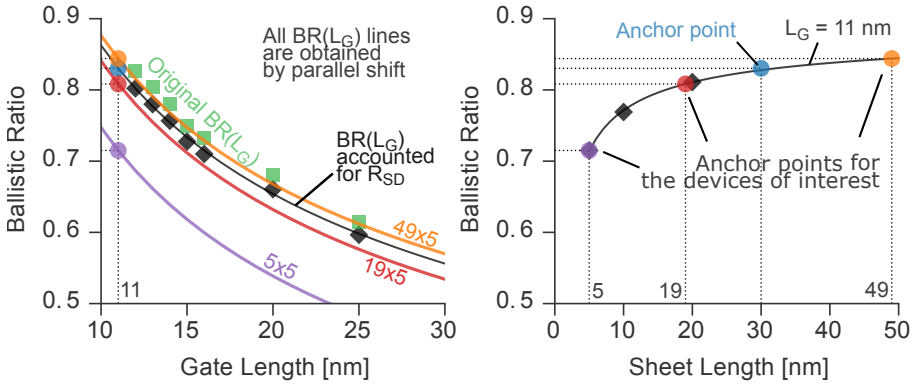


Figure 4.12: Explanation on the way how the ballistic ratio values were computed for device of interest.

4.2.2 Ballistic Ratio

Ballistic current estimation is only the first step in the device comparison. Ballistic current sets the ultimate performance limit for a device. However, we should estimate the actual saturation current which can be expected from a device if scattering is taken into account. As we explained in section 2.2.2, we do this through the BR. BR is a ratio of the actual current to the full ballistic current. Its estimation is a difficult problem as various scattering models should be included into the simulations: acoustic and inter-valley phonon scattering, surface roughness scattering, *etc.* Typically, the quasi-ballistic current may be accurately predicted through the Monte Carlo simulations [14], [45], [57]. These simulations were widely used for both planar devices and FinFETs. As an input, Monte Carlo simulation needs models of the device band structure [89]. The bulk silicon band structure is quite well known, so Monte Carlo may be applied "out of the box". For the FinFETs, some tricks related to the quantum correction have to be implemented on top of the simple simulation [90]. For the NWFETs, the quantum confinement gets too strong, and subband structure along the channel should be extracted by solving the coupled Schrödinger–Poisson equations [91]. This requires usage of the sophisticated simulators, not easily available. The development of such a simulator is out of scope of this work. Therefore, we rely on the already published data for the ballistic ratio as an input. Fig. 4.11 provides an overview of the available data.

These data were computed with various techniques. BR for FinFETs was computed with Monte Carlo which means that it accounts for the access resistance no matter if scattering in the channel is switch on or off. This is

because the full device structure with all the S/D features is simulated. BR for all the other structures was computed by solving the Schrödinger equation on two dimensional slices along the channel and then the one-dimensional Boltzmann transport equation along the channel [14]. With this approach the access resistance from the S/D regions is not accounted for. Ideally BR = $R_{ch.bal}/R_{ch.scat}$, where $R_{ch.bal}$ is a channel resistance in ballistic limit, and $R_{ch.sat}$ is a channel resistance when scattering is considered in the model. However, due to the presence of access resistance (R_{SD}), the formulas change in the following way:

$$\mathrm{BR(MC)} = \frac{R_{ch.bal} + R_{SD}}{R_{ch.scat} + R_{SD}}, \tag{4.1}$$

$$\mathrm{BR(BTE)} = \frac{R_{ch.bal}}{R_{ch.scat}}, \tag{4.2}$$

where BR(MC) is a BR coming out of the Monte Carlo simulations, and BR(BTE) is a BR coming out of the other approach based on the Boltzmann transport equation. Practically, because of R_{SD}, BR(MC) is about 3% higher than BR(BTE).

Coming back to the device layout (Fig. 4.2), it is clear that we should exploit three options: device made of three NWs, device made of two slightly elongated (up to 19 nm) NShs or device made of a single 49 nm long NSh. Each device, either NW- or NSh-based may be implemented with various body thicknesses. We will focus on two scenarios: 5 nm and 7 nm body thickness. Smaller diameters result in too poor BR values and, consequently, drive, while larger diameters penalize the footprint. All of the above mentioned options have nearly the same footprint, which ease the PPAC comparison.

In order to obtain BR values for all of these devices some manipulations with the data from Fig. 4.11 were made. We started with lowering BR of FinFETs by 3% to account for the impact of the R_{SD}. It is fair to assume that the BR between these NShs-based FETs and FinFETs is quite similar. Therefore, we first extracted BR for a FinFET having 11 nm gate length (blue point in Fig. 4.12a). This point was added to the three available points from the literature for the NSh device and was fitted with an analytical equation in the form of $a/L^b + c$, where L is a NSh length; a, b, and c are the fitting coefficients. This type of equation was chosen because in this case the BR saturates at large lengths at the reasonable value of about 0.89. Thus, with this equation we extrapolated BR for 49 nm long NSh (orange point in Fig. 4.12b). Next, we assumed that the function describing BR(L_G) dependency does not change for various dimensions. Thus, we made a parallel shift of the original function derived for a 30 nm FinFET. This is shown in Fig. 4.12a: purple is for the 5 nm

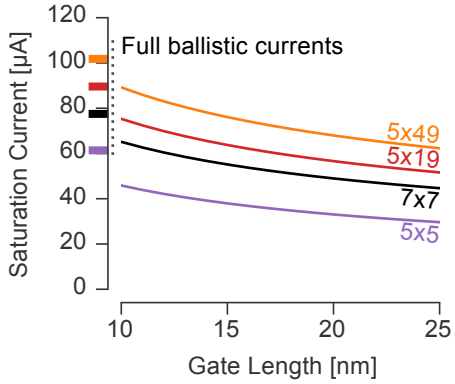


Figure 4.13: nMOS quasi-ballistic current in case of an ideal electrostatics at $I_{OFF} = 10\,\text{nA}$ and $V_{DD} = 0.6\,\text{V}$.

long NSh (essentially, this is similar to a NW), red is for the 19 nm long NSh, and orange is for the 49 nm long NSh.

Based on the information regarding BR, we can define saturation current targets per each options as shown in Fig. 4.13. There are no data on BR available for the 7 nm wide NWs. Therefore, we assumed the same BR(L_G) dependency as for the 19 nm×5 nm NSh. Because of the higher surface scattering impact on the 5 nm wide NWFET device, its initial benefits coming from higher injection velocity and higher number of carriers per effective width get diminished. Notice, that we have not reported current for the 7 nm wide NShs-based devices. The reason is related to their poor electrostatic control of the channel as will be explained in the next subsection.

4.2.3 Electrostatics

Up until this moment, we have discussed the quasi-ballistic transport, but we have not talked about the electrostatics difference between different device structures. This information may be reliably extracted using the conventional drift-diffusion model with density gradient corrections with Synopsys Sentaurus simulator [92]. Figure 4.14 summarizes the simulation results for various devices. 7 nm wide NW has relatively good electrostatics, however once we depart from the NW structure towards a 7 nm wide NSh, both subthreshold slope (SS) and drain-induced barrier lowering (DIBL) degrade severely (see the case of $19 \times 7\,\text{nm}^2$ NSh in Fig. 4.14). Decent electrostatics control can only be achieved with relaxed gate lengths, but in this case the device performance is penalized

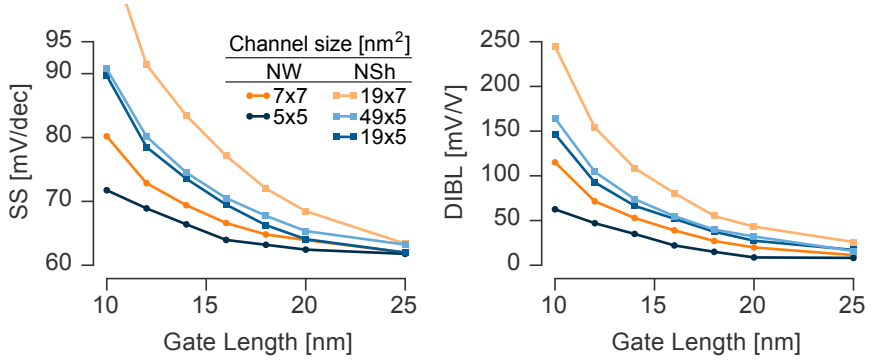


Figure 4.14: TCAD-predicted SS for NWs is the best. Yet, the NSh with the 7 nm thick body quickly loses electrostatics control with its width increase, thus the $19 \times 7\,\mathrm{nm}^2$ NSh is excluded from the analysis as well as the $49 \times 7\,\mathrm{nm}^2$ NSh (not shown).

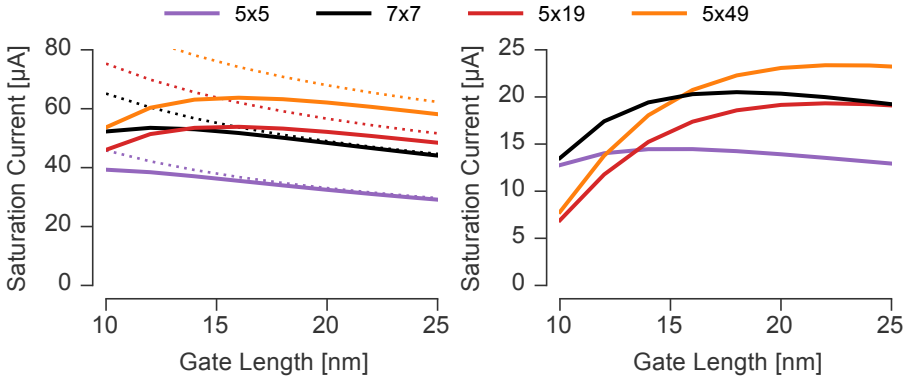


Figure 4.15: nMOS quasi-ballistic current with realistic electrostatics at $V_{DD} = 0.6\,\mathrm{V}$ and $I_{OFF} = 10\,\mathrm{nA}$ (left) or $I_{OFF} = 10\,\mathrm{pA}$ (right). No access resistance is considered. The lower off-current target is, the more important it is to maintain good electrostatics control (good SS). Dashed lines correspond to the case of nMOS quasi-ballistic current with an ideal electrostatics. Deviation from the ideal case is more pronounced for the short gate lengths.

through high channel capacitance and somewhat poor BR. Therefore, we exclude 7 nm wide NShs from our analysis.

With the gate length scaling, the DC performance would first increase because of increase in BR (Fig. 4.13), but at certain moment it would degrade as too

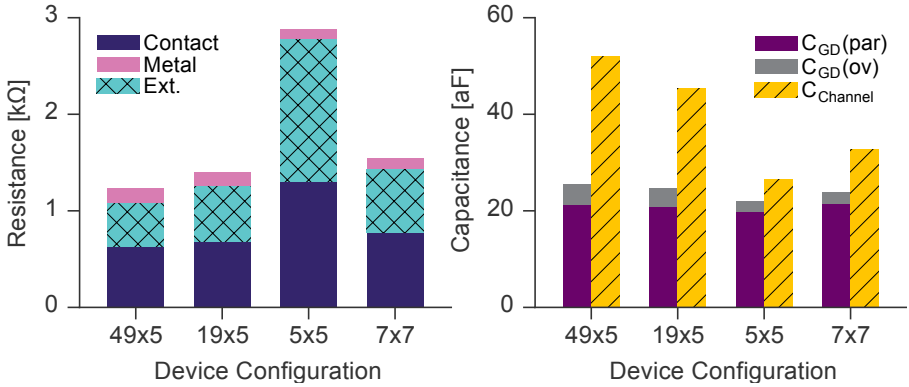


Figure 4.16: Parasitic resistances (S+D) and capacitances (Miller, channel) for pMOS devices. $L_G = 18\,\text{nm}$, S/D spacers are 10 nm thick with permittivity of 4.4, $\rho_C = 5{\times}10^{-9}\,\Omega\,\text{cm}^2$, extension doping is $3{\times}10^{20}\,\text{cm}^{-3}$. Drain is at the bottom.

aggressive gate length scaling would result in too high SS. Thus, there is always an optimal gate length. However, its exact value depends on the off-current targets and device structure. NW channel provides the best electrostatics control allowing to scale the gate length most aggressively. This is especially important for the applications with the low off-current targets (or, in other words, for high-V_{TH} case) as shown in Fig. 4.15. It is hard to imagine the vertical device technological process flow which would enable different gate lengths on a chip. Thus, a single gate length should be chosen as a trade-off between high performance (low-V_{TH}) and low power (high-V_{TH}) flavours.

4.2.4 Parasitics

Yet, before we fix the gate length, we should add parasitics into the picture. These parasitics vary for different devices. The DC performance is only affected by the resistance. The break-up of parasitics resistance for various devices is given in Fig. 4.16a. Clearly, the two major components are the extension resistance and the contact resistance. The model of the contact resistance was explained in details in the section 2.3.1. The extension resistance is determined by:

- the cross-section of the extension;
- the length of the extension;
- the extension resistivity, which is tightly linked to the doping level.

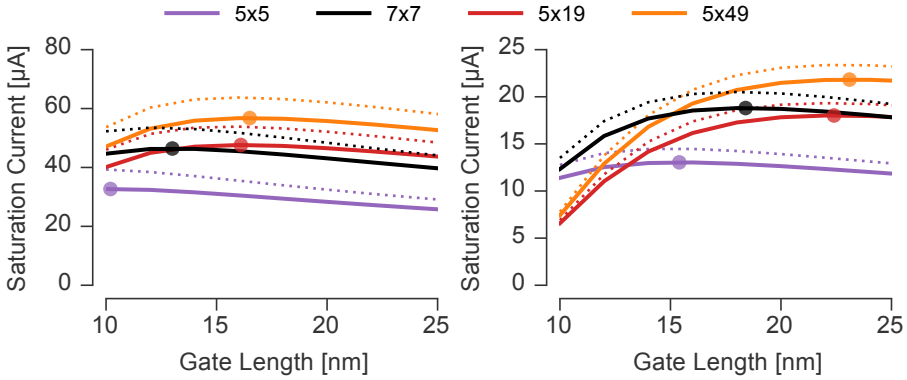


Figure 4.17: nMOS quasi-ballistic current with realistic electrostatics at $V_{DD} = 0.6\,\mathrm{V}$ and access resistance. Results are for low-V_{TH}, $I_{OFF} = 10\,\mathrm{nA}$ (left) and high-V_{TH}, $I_{OFF} = 10\,\mathrm{pA}$ (right) flavours. Dots indicate the optimal gate length per flavour per structure. The overall optimal gate length per structure is defined as an average of the best low-V_{TH} and the best high-V_{TH} L_G. Dashed lines correspond to the case of quasi-ballistic current with realistic electrostatics but without any access resistance.

The contact resistance is mainly defined by:

- the contacting surface area;
- the specific contact resistivity.

Let us start with the contact resistance. We assumed the specific contact resistivity to be $5{\times}10^{-9}\,\Omega\,\mathrm{cm}^2$ [93]. It is possible to go lower in this value [21], [94], but this requires extremely high doping levels in a semiconductor (in the order of $1{\times}10^{21}\,\mathrm{cm}^{-3}$), which might be hard to achieve with the in-situ doped epitaxially grown structures. As metallic electrode wraps the semiconductor, the contacting surface area is defined by the thickness of the metal electrode and the channel perimeter. The thickness of the electrode is kept fixed: 24 nm or metal pitch. The channel perimeter varies from one device structure to another one. So does the cross-section of the extension. As NWs are much smaller than NShs, they are penalized more. In order to compensate for high extension resistance, its length might be reduced or its doping level might be increased. The latter has already been discussed, it is hard to achieve too high doping levels. However, the length of the extension is an interesting parameter. In fact, it goes together with the parasitic capacitance. The smaller the extension, the higher the capacitance is. Thus, there is an *RC*-tradeoff which can only be visible when the *AC* performance is analyzed. Neverless, fixed S/D extension length (or *S/D spacer thickness*) does not prevent us from determining the

optimal gate length. Figure 4.17 indicates the impact of access resistance on the saturation current. By combining both high-performance and low-power flavours, we may obtain these numbers for the optimal gate length:

- **5x5 NWs** — $L_G \approx 13\,\text{nm}$;
- **7x7 NWs** — $L_G \approx 16\,\text{nm}$;
- **5x19 NShs** — $L_G \approx 19\,\text{nm}$;
- **5x49 NSh** — $L_G \approx 20\,\text{nm}$;

4.3 Performance on a ring oscillator level

4.3.1 Goals of the analysis

Technology benchmarking is often done on the ring oscillator level. It is particularly useful for the early technology assessment as it is relatively easy to run and to interpret the results afterwards. On top of this, the ring oscillator testbench may be used for the device optimization. In here, we demonstrate the way the S/D extension length may be optimized.

As for the S/D extension length optimization, it is, in fact, an *RC*-optimization problem. The smaller the extension is, the closer the S/D electrodes come to the gate electrode, which results in higher parasitic capacitance. Higher capacitance negatively impacts not only the operating frequency, but also the device power consumption. The only reason to make extensions smaller is to boost current by lowering access resistance which is directly proportional to the extension length. High currents are required to drive some load. This might be a load related to the fan-out of a logic gate output (the number of gate inputs it can feed or connect to). But it might also be a load related to the wiring between the logic gates.

Let us start with the capacitances related to a device. Fig. 4.16b provides an overview of capacitance break up for different devices at fixed S/D extension length (10 nm). Despite changes in the device channel structure, parasitic capacitances do not change much as they are mainly defined by the electrode geometry rather than by the channel geometry. In fact, these findings are consistent with the FinFET to NWFET transition. The parasitic capacitance of those two structures are very similar if the same fin height is maintained ([71]). It is the channel capacitance, which changes with the transition from one device structure to another. This may be easily perceived by the changes in effective width, although the true situation is a bit more difficult due to the quantum confinement effects as discussed in the previous section.

Table 4.1: BEOL R and C-values per micron wire length. Due to high resistance, the lowest RC-product is for the wire width increased beyond half pitch. With wire scaling, RC-delay goes up. The data are based on the internal imec measurements and simulations.

Wire material	Cu with TaN/Ru barrier				Cu with Mn/Ru barrier			
Metal pitch (nm)	32				24			
Wire width (nm)	16	18	20	22	12	14	16	18
R ($\Omega/\mu m$)	445	358	297	251	702	548	443	368
C ($fF/\mu m$)	0.19	0.21	0.24	0.27	0.20	0.23	0.27	0.34
RC ($fs/\mu m^2$)	86	77	71	69	138	123	118	125

Next, the interconnect related load. With scaling, a cross-section of metal wires gets smaller and smaller together with a distance between them (Table 4.1). This results in a high wire RC-delay, comparable to a gate delay, especially for the bottom-most (and, thus, the most dense) metal layers [95].

In case the performance is poor due to a high fan-out number, shrinking of S/D extension is not very effective because load increases simultaneously with an increase of drive current. The net result is thus determined by the relative pace of these two competitive processes. However, in case the performance is limited by the interconnects, shorter S/D extensions might help. In fact, this is all about front end of line (FEOL) and BEOL balancing. To do this balancing exercise correctly, we need to obtain proper values for both fan-out and typical wire length distribution. This requires to go through the whole system on chip (SoC) design loop including the place and route phase.

4.3.2 Benchmark description

Figure 4.18 shows the M_X wire length and fan-out distributions for the OpenCores LDPC decoder [96] critical path. This design is quite a good representative of a graphics processing unit (GPU) / digital signal processor (DSP). Both wire length and fan-out distributions are very skewed with heavy tails. Therefore, it is better to look at the median value rather than at the mean value. The median value for the wire length is 65 CGPs long wire. The median value for the fan-outs is three. The LDPC design is actually quite heavily BEOL-loaded, therefore it would be fair to drop a typical wire length a bit. We use 50 CGPs long wires for the baseline load.

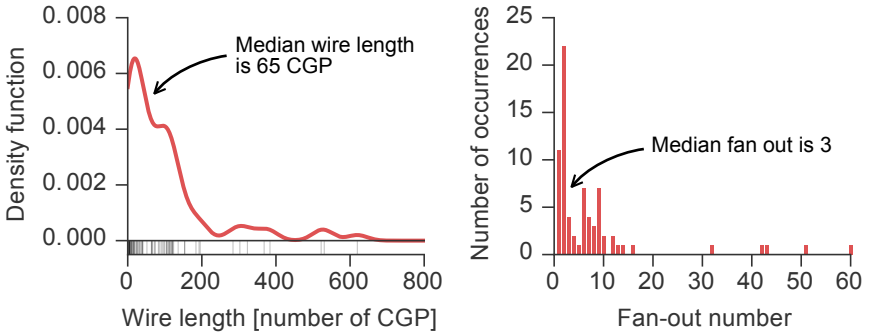


Figure 4.18: Wire length and fan-out distribution comping from the OpenCores low-density parity-check (LDPC) decoder [96] critical path data.

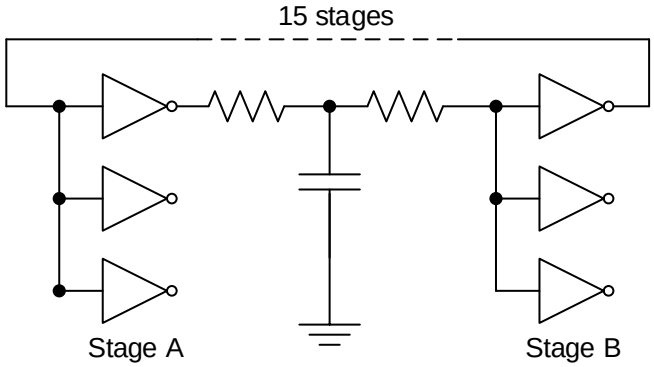


Figure 4.19: Schematic of the ring oscillator testbench.

Ring oscillators used for benchmarking may be made of simple inverters (INV) [1], NAND gates [97], or of a mixture between INV, NAND and NOR gates [98]. Our baseline testbench (Fig. 4.19) consists of 15 stages ring oscillator, all of them are inverters. Each stage drives three INV gates (fan-out is three). The wiring between each stages is 50 CGPs long. Our analysis is focused on the 5 nm node, which implies tight ground rules (see discussion on these rules in chapter 3). As such, the M_X-level metal pitch we consider at the 5 nm node is 24 nm. The related *RC*-delay may be calculated based on the data from Table 4.1. In case of 50 CGPs long wires, the wire delay is 0.30 ps (our CGP is 32 nm).

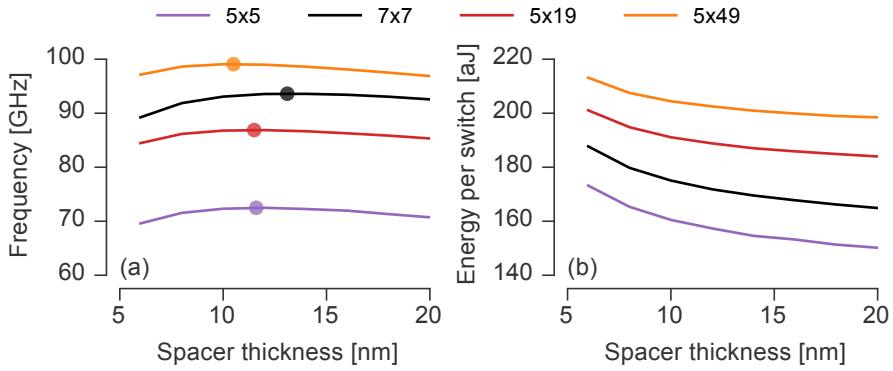


Figure 4.20: (a) There is always a S/D spacer thickness value corresponding to the maximum oscillation frequency. It depends on device structure and it is indicated with circle. (b) Energy per switch always increases with spacer thickness scaling because of higher device capacitance. In here, $V_{DD} = 0.6\,\mathrm{V}$ and $I_{OFF} = 10\,\mathrm{nA}$.

4.3.3 Simulation results

First, let us see how oscillation frequency (Fig. 4.20a) and energy consumption (Fig. 4.20b) change with variation of S/D extension length. As expected, energy consumption gets worse with S/D spacers scaling. In turn, there is always the S/D spacer thickness corresponding to the maximum oscillation frequency. Arguably, spacer thickness optimization is not very important for the device optimization as frequency is not very sensitive to spacer variations. Nevertheless, we continue with the S/D extension lengths corresponding to the maximum frequency unless other is specified:

- **5x5 NWs** — $T_{SP} \approx 11.6\,\mathrm{nm}$;
- **7x7 NWs** — $T_{SP} \approx 13.1\,\mathrm{nm}$;
- **5x19 NShs** — $T_{SP} \approx 11.4\,\mathrm{nm}$;
- **5x49 NSh** — $T_{SP} \approx 10.5\,\mathrm{nm}$;

Next, we see how the device performance changes with leakage targets variations to cross-check our gate length choice from the DC performance study. Fig. 4.21a clearly indicates that oscillation frequency drops with leakage reduction. There is no surprise here as off-current has been targeted with work function tuning. That said, for low off-current targets V_{TH} gets high which inevitably results in performance degradation. It is more interesting though that there are no cross-over points between different device options as the gate length for all of the devices was chosen as a trade-off between high-V_{TH} and low-V_{TH} flavours.

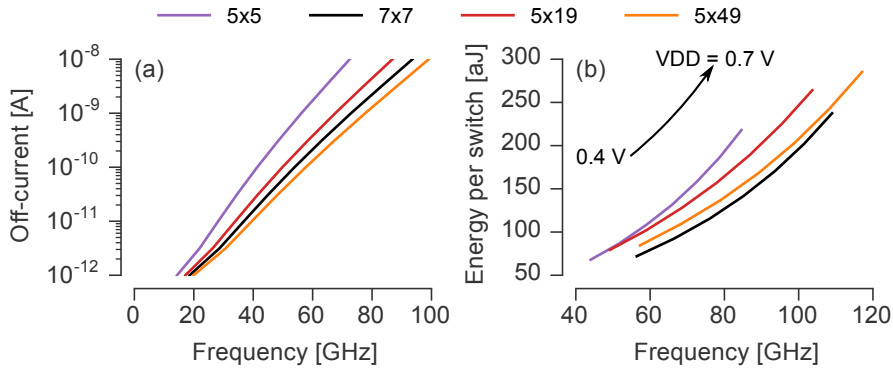


Figure 4.21: (a) There is no clear cross-over between different device options depending on off-current targets, as the gate length for all the devices was chosen to satisfy both high-V_{TH} and low-V_{TH} flavours. $V_{DD} = 0.6\,\mathrm{V}$. (b) At high V_{DD}, 49x5 NShFET is faster than 7x7 NWFET, but consumes more energy. Difference in speed decreases for smaller V_{DD} as NWFET has better electrostatics. 5x5 NWFET is not a competitor to any device options due to its low drive. $V_{DD} = 0.6\,\mathrm{V}$ and $I_{OFF} = 10\,\mathrm{nA}$.

Fig. 4.21b shows how the ring oscillator performance metrics change with V_{DD} variations. A couple of observations may be made. First, NWFETs are better suited for V_{DD} scaling than NShFETs. This is related to their better electrostatics control. Next, the 5x5 NWFET is an outsider loosing to all other device options. The 7x7 NWFET is a winner as it delivers highest performance at same energy consumption or, other way round, it provides least energy consumption for the same oscillation frequency. Last, despite the 7x7 NWFET looks the best, when it is compared to 49x5 NShFET, the latter is faster if run at same V_{DD} though more power hungry. This explains why 49x5 NShFET is ahead of 7x7 NWFET in the Fig. 4.21a.

We started our discussion on NSh *vs.* NW channels by saying that NSh should be better suited for high loads because of its higher drive. Fig. 4.22 confirms this hypothesis. Without any interconnect load the 7x7 NWFET provides the highest frequency, and the 5x5 NWFET is comparable to 19x5 NShFET. However, for high loads, drive requirements get more important. As such, for the 300 CGPs long wires, both NShFETs outperform both NWFETs.

Last, let us discuss how the optimization of S/D spacers changes if high doping levels in these spacers cannot be achieved. We consider two scenarious: a) spacer thickness is kept fixed (at its optimal value for the $3{\times}10^{20}\,\mathrm{cm}^{-3}$ doping level), b) spacer thickness thickness is getting re-optimized for each doping level.

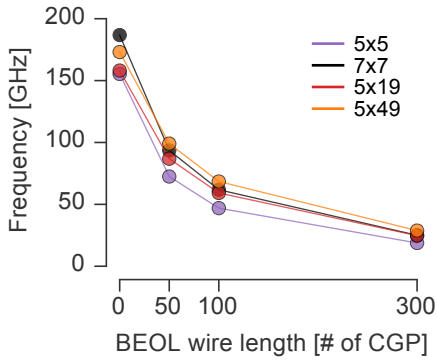


Figure 4.22: In case there is no BEOL wires in between the ring oscillator stages, the 7x7 NWFET provides the highest frequency. The 5x5 NWFET is comparable to the 19x5 NShFET. However, as soon as the load increases, drive requirements get more important and already for the 300 CGPs long wires both NShFETs outperform both NWFETs. $V_{DD} = 0.6\,\mathrm{V}$ and $I_{OFF} = 10\,\mathrm{nA}$.

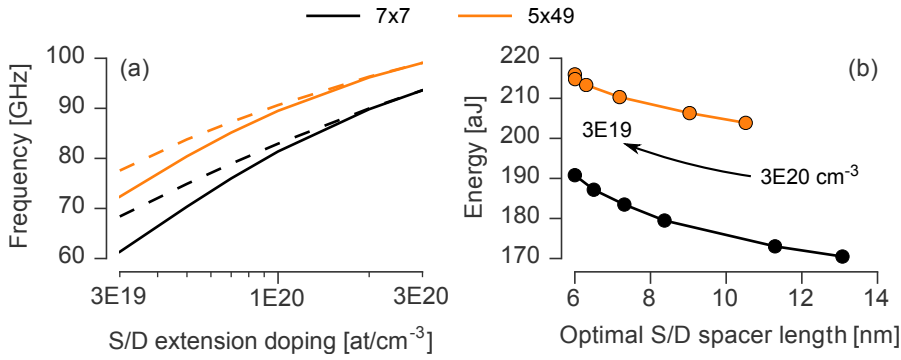


Figure 4.23: (a) In case the S/D extensions are not highly doped, 7x7 NWFET is 15% slower than 49x7 NShFET if same spacer is used (solid lines). Spacers may be thinned to boost oscillation frequency (dashed lines), however this comes at the expense of increased energy (b). $V_{DD} = 0.6\,\mathrm{V}$ and $I_{OFF} = 10\,\mathrm{nA}$.

The first scenario is shown with solid lines and the second one — with dashed lines (see Fig. 4.23a). We may see that S/D spacer thickness re-optimization may result in quite significant performance gains. In addition, because of the smaller extension cross-section for the 7x7 NWFET case than for the 49x5 NShFET, lower doping levels penalize NWFET performance more than NShFET

performance: NWFET is 5.5% slower for the $3{\times}10^{20}\,\mathrm{cm}^{-3}$ doping and 15% slower for the $3{\times}10^{19}\,\mathrm{cm}^{-3}$ doping. Spacer re-optimization may lower the difference down to 12%. However, this spacer re-optimization happens at the cost of higher energy consumption as shown in Fig. 4.23b.

4.4 Summary and conclusions

In this chapter, we focused on two aspects. First, we wanted to choose a vertical device, to understand whether it should be made of several NWs or of a single nanosheet. Second, we tried to introduce to the reader our benchmarking methodology which covers various abstraction levels. We started from the discussion on the impact of quantum confinement on device performance and finished with the impact of the interconnect load. By doing this, we emphasized that certain phenomena may play different role in the overall performance than they appear to play. For example, because of quantum confinement, NWs have higher current per effective width than nanosheets. However, these benefits get screened by access resistance. On top that, the effective width of several NWs is still smaller than that of a single nanosheet, having similar-footprint. Therefore, a nanosheet has a higher absolute drive and it is better capable of driving long interconnects.

We also introduced the test vehicle which is supposed to represent a critical path of mobile SoC. *Based on this test vehicle*, we conclude that a VFET made of several NWs is better than a nanosheet-based device. However, not every NW is suitable. As such, we demonstrated that device based on 5 nm diameter NWs cannot provide sufficient drive and, therefore, the we choose a VFET made out of three 7 nm diameter NWs as the best vertical device made with 5 nm ground rules. It is this device, which we will use further on for benchmarking with lateral devices and vertical tunnel FETs.

The key messages of the chapters are listed below.

- Holistic approach to the benchmarking is essential. It allows to identify the regimes where one device outperforms another making the conclusions assumptions specific rather than just black and white.
- Based on our test vehicle, which is a simplified representation of a high-performance mobile SoC, NWs-based device outperforms nanosheet-based device.
- Other factors like variability may affect the derived conclusion and require further study. Yet, as the main source of variability is related to metal grains in the gate stack [99], we may already intuitively predict that NWs will be more resistant to this variability than nanosheets. As the

channel is rather narrow, any work function fluctuation on our side of the channel will be balanced by the similar fluctuations on the other side of the channel. A nanosheet has only one pair of channel sidewalls (elongated ones) where this effect takes place, while for a NW, this effect is pronounced throughout its perimeter.

CHAPTER 5

Vertical Tunnel FET

5.1 Why TFET?

In the previous chapters we have discussed the potential benefits of vertical architecture. Probably, its key advantage is in the freedom of gate length and spacer thickness optimization. Ultimately, this may lead to the reduced power consumption with respect to lateral architecture, while performance is preserved. However, vertical architecture does not bring any extra room for the V_{DD} scaling. Yet, this is the most powerful knob in the reduction of energy consumption. The main reason for the poor V_{DD} scaling is related to the fundamental limit of MOSFETs: their SS cannot become steeper than 60 mV/dec at room temperature. This means that in order to maintain the performance (in the first order, it is defined by the gate overdrive, $V_{GS} - V_{TH}$), the V_{TH} should be scaled together with V_{DD} at the same pace. Unfortunately, this results in the excessive leakage which is intolerable, especially for the mobile applications.

Figure 5.1 illustrates the SS issue on the example of an nMOS device. Similar reasoning may be applied to the pMOS device. The doping in the source and drain regions is quite high which pushes the Fermi level into the conduction band. At the room temperature, there are *always* some electrons having energy above the Fermi level, as they follow the Fermi-Dirac distribution. Therefore, when the gate barrier is lowered with the gate voltage, current *gradually* increases as more and more carriers form the Fermi-Dirac distribution tail contribute to the current flow.

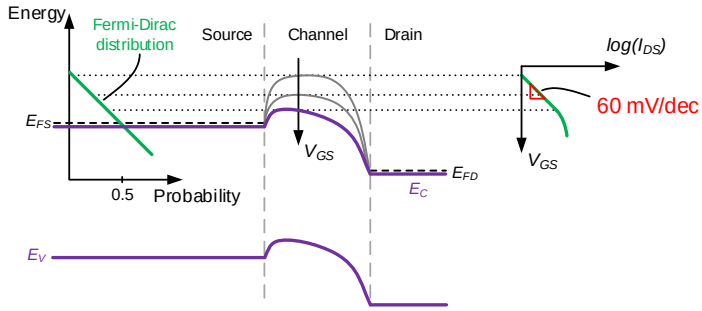


Figure 5.1: Operating principle of a MOSFET sets the limit on subthreshold slope: 60 mV/dec at room temperature.

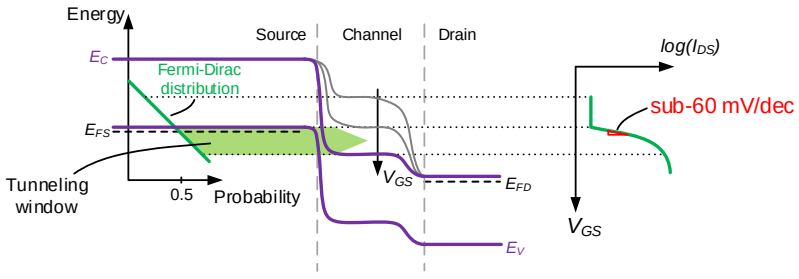


Figure 5.2: As TFET operating principle relies on the tunneling mechanisms rather than the thermal injection, its subthreshold slope can reach values smaller than 60 mV/dec.

The operating principle of TFETs is different. Figure 5.2 illustrates the way an n-TFET works. Unlike in MOSFETs, the source here is p-doped with the Fermi level around the valence band energy. This means that the tail of the Fermi-Dirac distribution is in the forbidden gap and does not contribute to the current flow. When the sufficiently high gate voltage is applied (so that the conduction band energy of the channel is below the valence band of the source), the interband tunneling can occur within the so-called tunneling window: device turns on with a sub-60 mV/dec slope.

Therefore, because of the steeper slope in TFETs, V_{DD} may be reduced while leakage and overdrive remain unaffected. Yet, the drive current which comes out of a MOSFET is different from the current of a TFET even at the same overdrive! Being more specific, a TFET current is *smaller*. Therefore, the direct

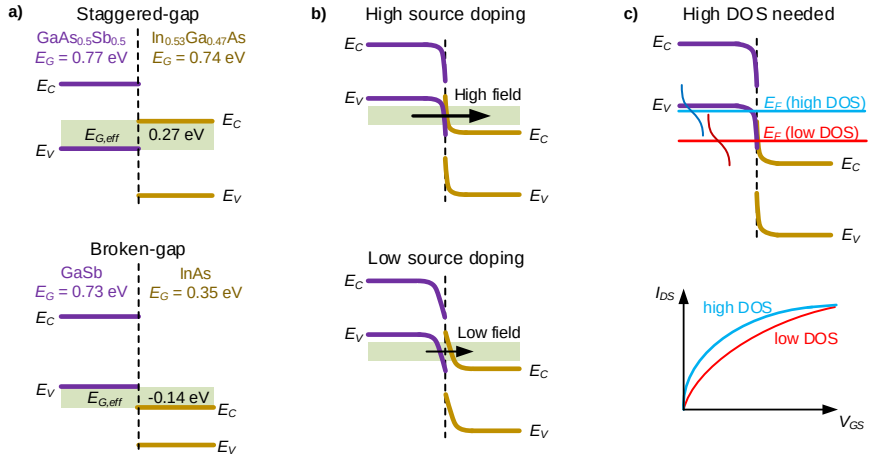


Figure 5.3: TFET Requirements.

comparison is not straightforward. In this chapter we will go through different aspects of TFET design and we will try to understand what their impact on device performance is. Once we have this information, we will benchmark TFETs with MOSFETs in the next chapter.

5.2 TFET device design and simulation

Essentially, the TFET device is a gated p-i-n diode. However, in order to achieve good performance the device configuration should be well optimized. Let us try to see, what are the major general device requirements. A nice summary is given in [100], we will reproduce it here. First, the heterostructure is needed: source material should differ from channel and drain material. Energy should be aligned as shown in Fig. 5.3a leading to either a staggered-gap device or a broken-gap device. In both cases the idea is to reduce the effective bandgap at the source-channel junction to boost band-to-band tunneling (BTBT) current, and to keep large bangap at the drain side to effectively suppress ambipolar BTBT currents. Next, the doping on the source side should be very high in order to increase the electric field and, thus, current (Fig. 5.3b). With high source doping, the depletion length on the source side gets shorter, the bands bending gets steeper. Last, but not least, the density of states (DOS) in valence band (for n-TFET) and conduction band (for p-TFET) should be large to avoid too much of degeneracy at the source and boost device performance (Fig. 5.3c).

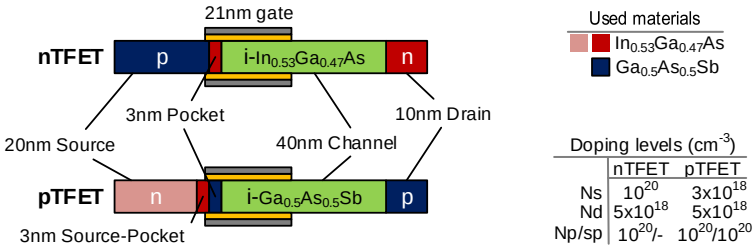


Figure 5.4: TFET devices with 10 nm thick body. No S/D contacts are shown. These contacts are 20 nm long and have high ($1\times10^{20}\,\mathrm{cm}^{-3}$) doping in order to lower specific contact resistivity.

Based on these principles, well optimized devices were designed [100]. These TFETs (Fig. 5.4) have 3 nm pockets in the source which boosts drive preserving steep SS [101], with an extra source pocket for the p-TFET [102]. Both devices are III-V based as they are predicted to deliver best performance [35], [100]. Devices are strongly asymmetrical, with a 2× longer channel than the gate to suppress ambipolar current and to reduce the gate to drain capacitance. In order to fabricate these complex heterojunction structures with well defined pockets, the in-situ doped epitaxial growth seems to be the only feasible solution. Which implies that these devices have to vertical.

In order to simulate these TFETs, we used the in-house developed 2D quantum mechanical simulator [103]. It relies on the continuum approach, based on a 15-band envelope function formalism. Moreover, the models of the simulator were calibrated to the diode measurements [104]. The limitation of the simulator is that it treats transport as fully ballistic. Due to the absence of scattering models the drive current at high bias is likely overestimated. However, as III-V materials, which we are talking about here, have lighter effective mass than silicon, we can expect the ballistic ratio to be at least comparable to that of MOSFETs (more than 75%). Therefore, this overestimation should not be too high. Another issue is related to the fact that the simulations are performed in 2D. Because of this, we only explore NSh channels, which are 49 nm long and 10 nm thick. The footprint of such a device is similar to a footprint of a VFET, which we looked at in the previous chapter.

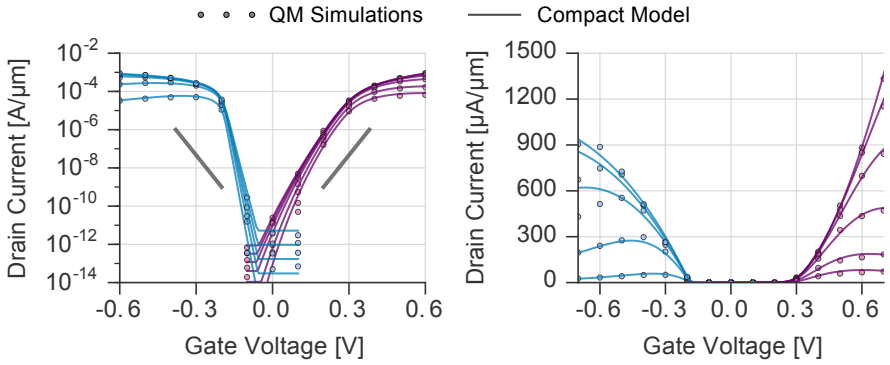


Figure 5.5: Compact model is flexible enough to fit the reference currents predicted by the quantum mechanical (QM) simulator. Fitting results for transfer characteristics at different drain biases are demonstrated here with current plotted in the logarithmic scale (left) and linear scale (right).

5.3 TFET compact model

5.3.1 Choice of a compact model

The quantum mechanical simulator is not applicable for the circuits analysis, as a simulation of just a single device takes several hours. The analytical compact model is required. TFETs are emerging devices, which means there is no consensus in the industry on how the device should look like. Because of this, there is no compact model, which would be an industrial standard like BSIM-CMG for MOSFETs. There are various analytical models available in literature [105], [106], [107], [108], each of them dedicated to a specific device structure (*e.g.* homojunction Si-based TFET) or/and focusing on modeling of particular effects. Following our reasoning in section 2.1.2, we have been looking for a versatile TFET analytical model, which would capture the essential physics and would be easily tunable. Finally, we have made a choice towards the model from the Notre Dame and Udine Universities [109].

5.3.2 Current fitting

As we demonstrate it in Fig. 5.5, this model may be well fitted to the reference currents predicted by the quantum mechanical simulator. However, in order to achieve this, we improved the original model by the introduction of DIBL-

like effect: V_{TH} and SS dependency on the drain bias. We made n and V_{TH} parameters dependent on V_{DS}:

$$n = n_0 + nSS \cdot V_{DS}, \tag{5.1}$$

$$V_{TH} = V_{TH0} + DVTP0 \cdot V_{DS}^{DVTP1}, \tag{5.2}$$

where n and V_{TH} have the same meaning as it is in the original manuscript, n_0 is the parameter defining SS at $V_{DS} = 0$, while nSS is a fitting parameter which alters SS with V_{DS}. V_{TH0} is a parameter defining V_{TH} at $V_{DS} = 0$, and $DVTP0$ together with $DVTP1$ are fitting parameter similar to BSIM-CMG model. The compact model fitting was done with the non-linear least-square minimization algorithm [110].

5.3.3 Charge fitting

Along with the current predictions, intrinsic device charge has to be predicted as well. It is common in literature to base the analytical models on capacitances, however this approach is prone to convergence issues and/or errors [43]. One of the possible errors is related to the charge conservation. According to [111], the best way to guarantee charge conservation is to formulate the model to be charge based. We followed these recommendations, although we had to use our own analytical equations for charge calculation, as the Notre Dame model [109] works with capacitances. The developed set of equations allows flexible fitting. However, it lacks any physics insights, which makes the fitting procedure somewhat cumbersome. Source charge is computed in a very easy way with Eq. 5.3, while drain charge computation is a bit more difficult and it is based on the combination of two functions: an asymmetrical generalized logistic function (Eq. 5.4 [112]) and a linear function (Eq. 5.5). Smooth transition between the two functions is achieved with Eq. 5.6.

$$Q_S = \exp(A \cdot V_{GS}) \cdot B, \tag{5.3}$$

where A and B linearly depend on V_{DS}.

$$\begin{aligned} X_b &= X_{50} + (1/m) \cdot \ln\left(2^{1/s} - 1\right) \\ \text{Num} &= T - B \\ \text{Den} &= \left(1 + 10^{m \cdot (\mathrm{X_b} - V_{GS})}\right)^s \\ Q_{Dlog} &= B + (\text{Num} \,/\, \text{Den}) \end{aligned} \tag{5.4}$$

where T and B define the plateaus at the left and right ends of the curve. X_{50} defines the V_{GS} where Q_{Dlog} function is in the middle between T and B values. X_b defines the inflection point and s allows for asymmetry.

$$Q_{Dlin} = k \cdot (V_{GS} - V_{DS} + V_T), \tag{5.5}$$

where k is a proportionality factor and V_T may be interpreted as some kind of a threshold voltage.

$$Q_D = Q_{Dlog} + (1 + \tanh(K \cdot (V_{GS} - V_{DS}))) \cdot (Q_{Dlin} - Q_{Dlog})/2, \tag{5.6}$$

where K defines the rate of transition between the functions Q_{Dlog} and Q_{Dlin}. At this point, we should comment why we had to use the extra Q_{Dlin} function. In the case of a logistic function alone, the drain charge at hight V_{GS} bias would be constant and given by the T value. As C_{GD} is a derivative of charge, it would become zero. This is not possible, in fact C_{GD} should saturate to a certain value [113], [114], which means that the drain charge should keep on increasing *linearly* with V_{GS}.

Figure 5.6 illustrates the charge fitting results as well as derived capacitance values. It might look like the fitting is of a poor quality for pTFET, however the root cause of the discrepancy is related to limitations of the QM simulator: the Q-V values are not based on a self-consistent simulation (potential profile is semiclassical, charge calculation is quantum-mechanical), such that some approximations have to be used in the charge extraction. The noisiness of the Q-V data originates from the presence/absence of resonances in the system, and should not be taken quantitative. This is especially true for the pTFET device, where source has a complex double-pocketed structure.

Another impact of having a heterojunction pocketed device with optimized source and drain doping levels is related to the relative values of C_{GS} and C_{GD}. Typically, the C_{GD} value in TFETs is believed to be high. It resulted in a number of paper warning about the enhanced Miller effect in TFETs [113], [115], [116], [117]. Yet, we demonstrate in Fig. 5.6 that the C_{GD} value has the same or even smaller value than C_{GS}, and, thus the overall gate capacitance is defined by *both* capacitances across a wide V_{GS} range. In fact, these findings are consistent with the results from Intel [118], which makes us believe that our findings are correct.

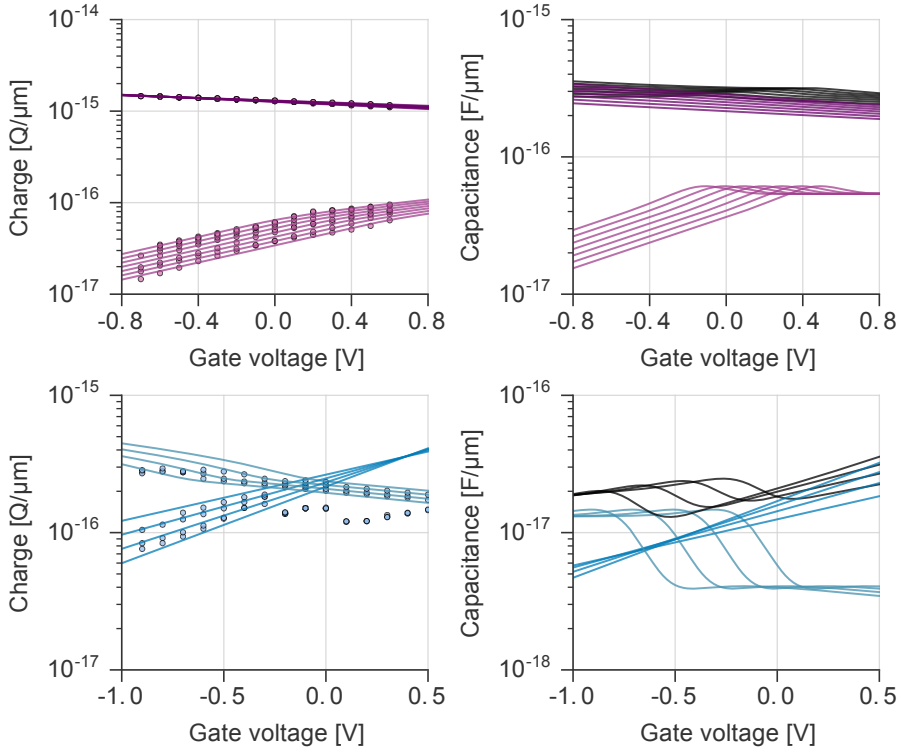


Figure 5.6: The developed set of analytical equations allows us to accurately capture charge predicted by the quantum mechanical simulator in a compact model (left column). The capacitances are derived afterwards automatically (right column). The top row is for the nTFET, the bottom row is for the pTFET.

5.4 Parasitics

5.4.1 Parasitics modelling

Being able to fit current and charge of the intrinsic device is an essential feature of a compact model, but not the only one. In addition, it should be capable of estimating parasitics. As the device is vertical, we reuse the same parasitics macro model which we used for VFETs, but with different material properties (III-V over silicon). Yet, there is one more difference. It is related to the gate to drain capacitance. As the TFET gate length is much shorter than the channel

Table 5.1: TFET parasitics. Doping underneath S/D contacts is $1{\times}10^{20}\,\mathrm{cm}^{-3}$ to reach $\rho_C = 1{\times}10^{-8}\,\Omega\,\mathrm{cm}^2$. S/D spacers relative permittivity is 4.4.

	nTFET	pTFET
R_S [kΩ]	1.65	0.78
R_D [kΩ]	0.61	7.22
C_{GS} [aF]	17.9	14.0
C_{GD} [aF]	8.44	9.08
$C_{Channel}$ [aF]	12.7	16.3

length, the capacitance has to be computed a bit differently.

The TFET channel is undoped, thus we do not compute fringing capacitance from the gate electrode to the channel. Essentially, we use formula for k_{ext} from the Table 2.1 with parameter s equal to the length of the ungated part of the channel. Figure 2.22 illustrates the impact of s parameter value on the fringe capacitance; it is called the "undoped length" over there.

5.4.2 Parasitics break-up

TFETs are asymmetrical devices with a source being different from a drain. Not only the geometry is different (see Fig. 5.4), but also the materials and doping levels. This has a direct impact on access resistance. We tried to account for these differences in resistivity using both literature data [119], [120], [121] and in-house measurements on III-V resistivity *vs.* doping level. Contacts are treated differently. Independently of the material, we assume $1{\times}10^{-8}\,\Omega\,\mathrm{cm}^2$ specific contact resistivity. This value is not as good as it is for silicon, but we find it to be consistent with the data from literature [94], [122]. Still, in order to reach this contact resistivity value, we assumed that doping in the contact region is quite high ($1{\times}10^{20}\,\mathrm{cm}^{-3}$).

Table 5.1 summarizes values for parasitic resistances and capacitances per device and per S/D, as they are all different. Device electrode extensions follow the actual layouts like it was in the case of regular VFETs (see the previous chapter). The striking value is a drain access resistance of a pTFET. The reason behind is that a) the drain extension is not highly doped, and b) it is p-type doped. This results in a very high extension resistance. Yet, as currents from TFETs are not that high, the voltage drop on this resistance is not that prominent.

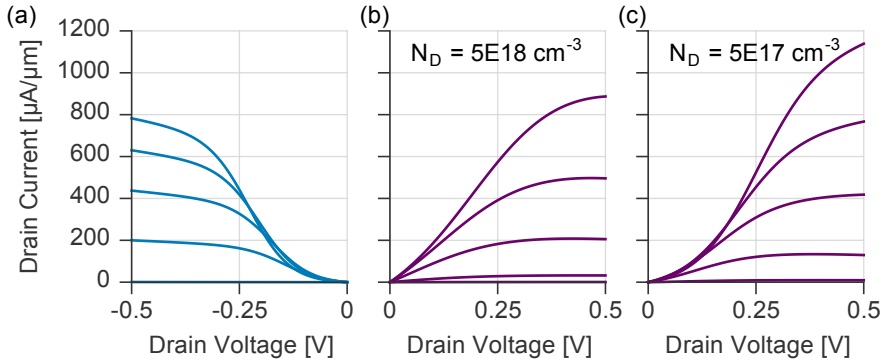


Figure 5.7: Output charactersitics of a) pTFET, b) nTFET with drain doping of $5{\times}10^{18}\,\mathrm{cm}^{-3}$, c) nTFET with drain doping of $5{\times}10^{17}\,\mathrm{cm}^{-3}$. The off-current for all the devices is aligned to $10\,\mathrm{pA/\mu m}$ at $V_{DS} = 0.5\,\mathrm{V}$. Gate voltage is varied from $0.2\,\mathrm{V}$ to $0.6\,\mathrm{V}$.

5.5 TFET output chacterisitcs

5.5.1 Delayed onset

Figure 5.5 demonstrates the input characteristics of our devices, however, it is quite interesting to look at the output characteristics as well (Fig. 5.7). Let us focus on the pTFET first. There are few peculiarities visible in the $I_{DS} - V_{DS}$ curves. First, current at low drain bias is very small. Second, for low drain bias current drops with increase of gate voltage. Last, the saturation of current happens at relatively high drain bias. All these effects are quite correlated, and there are several explanations available in literature: too large effective tunnelgap ($E_{G,\,eff}$) [123], too thick gate oxide [123], [124], too low source doping [123]. The high source doping helps not only to increase the tunneling probability, but it also ensures the availability of carriers [125], [126].

Similar observations may be made regarding the nTFET, although all the effects are less pronounced. In addition to the nTFET device described above, we look at a similar device but with the different drain doping: lowered to $5{\times}10^{17}\,\mathrm{cm}^{-3}$. In the second case the SS gets steeper, however the delayed onset in the output characteristics gets more visible, which essentially means lower current as small V_{DS}. We will come back to the impact of this later on.

At this moment, just by looking at all of these devices, we may get the impression that the steeper the slope is, the longer the $I_{DS} - V_{DS}$ onset is. Strictly speaking

this is not a generic statement, but it gives a clue that the co-optimization of both transfer and output characteristics is not that easy, yet essential. Poor output characteristics have a direct impact on circuits performance: rise and fall times increase drastically, moreover, inverter may not even be able to reach V_{DD} and V_{SS} rails [117], [127]. On top of that, the static noise margins (NMs) get degraded along with the inverter gain degradation [127]. In the next sections, we will try to understand the impact of these effects on the ring oscillator performance.

5.5.2 Noise margins

Typically NMs are discussed in the context of the static random access memory (SRAM), but we will look at the inverter voltage transfer characteristics (VTC) and extract static NMs from there. First, let us build an inverter like if we would have MOSFETs devices instead of TFETs. We target same off-current (10 pA) for both n- and p-devices with the gate work-function tuning and build an inverter out of these devices. Figure 5.8a shows a VTC of such an inverter. There are two clear observations. First, n- and p-TFET devices are unbalanced, which results in the low NM not being equal to high NM. Second, the V_{OUT} transition from V_{DD} to V_{SS} is not very sharp and steep which degrades both NMs and gain.

As we operate with a compact model, we have a freedom to artificially adjust some parameters of our devices. We focus on the output characteristics optimization and try not to change any other device parameters. All the changes are applied both to n- and p-TFETs, although we illustrate their impact on p-TFET characteristics only. Fig. 5.8b shows the case with slightly improved devices. Their output characteristics saturation happens earlier. In a compact model is was achieved by making the electric field at the source-channel interface independent on the drain-source bias (parameter $\gamma_1 = 0$). The impact on the VTC is minor. Fig. 5.8c illustrates the impact of the delayed onset by removing it completely (parameter $\lambda = 0$). The improvements are major. VTC is steep and it resembles MOSFET characteristics. The last figure (5.8d) combines the previous two improvements together. It is quite clear from this simple study that somehow the delayed onset should be suppressed in order to get good VTC. In addition to that the n- and p-TFET devices should be well balanced.

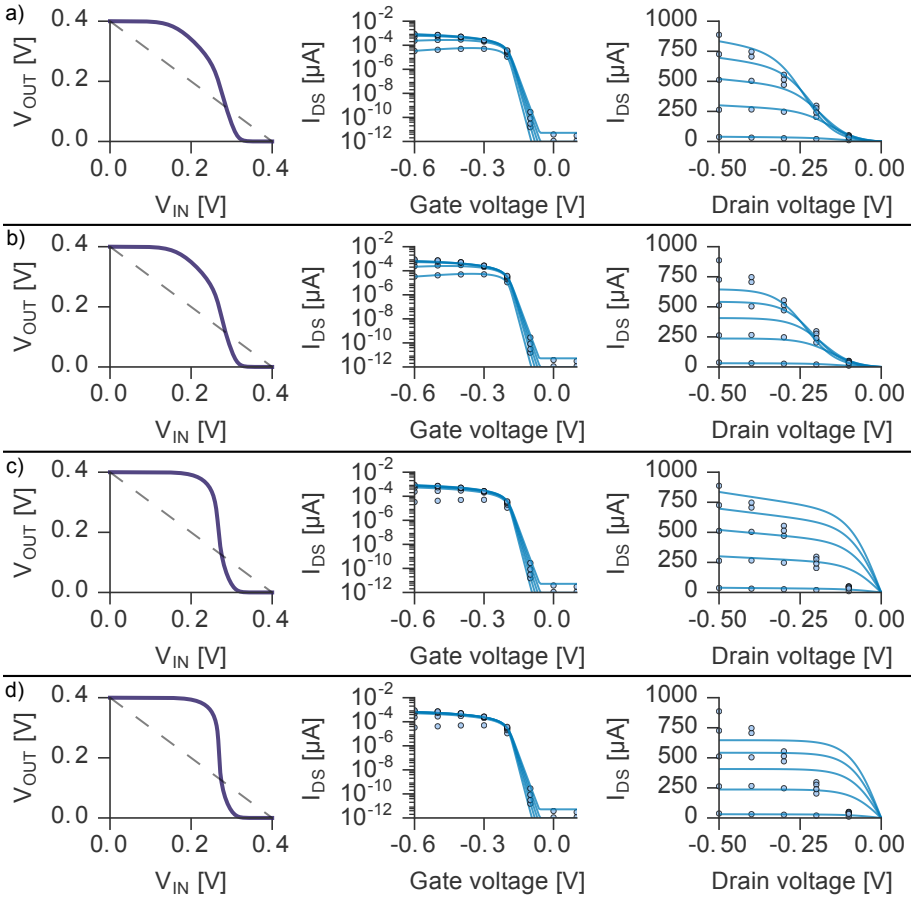


Figure 5.8: While artificially improving output characteristics of the TFET devices, we watch the impact of these improvements on VTCs. a) No improvements made. b) Output characteristics saturate earlier. c) The super-linear onset is suppressed. d) Both improvements are combined together.

5.5.3 Interplay with subthreshold slope

The VTC unbalance comes form the fact that the pTFET has about 2× steeper SS than the nTFET (Fig. 5.5). This makes the pTFET stronger than the nTFET and shifts the VTC to the right. It is possible to a) balance VTC; and b) study the impact of SS on NMs by artificially making SS of nTFET and pTFET equal in the compact model.

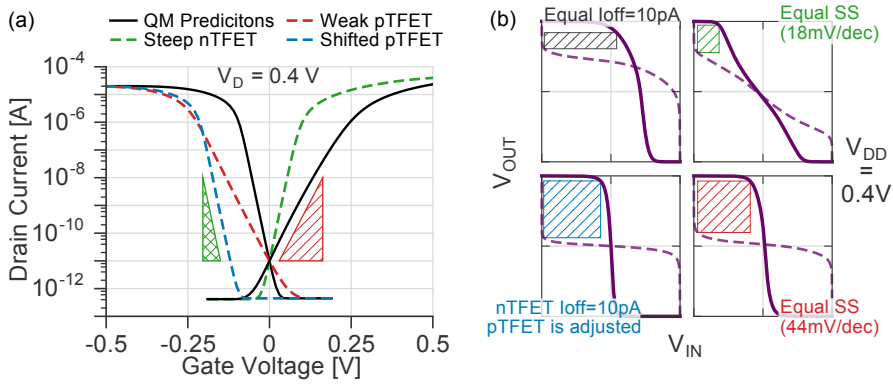


Figure 5.9: (a) TFET $I_{DS} - V_{GS}$ of QM simulated device and devices with artificially changed SS in the compact model to study impact of SS and pTFET IOFF on (b) inverter VTCs. Steep SS is not sufficient for good VTC.

Figure 5.9 illustrates this idea. Again we start with the devices as they are, but aligned to the same off-current. The VTC is poor and unbalanced. Next, we artificially make the SS of the nTFET as good as it is of the pTFET. We change nothing else in model cards. The VTC is now well balanced as expected, but NMs are drastically degraded! We can make the exercise in the different direction by degrading the SS of the pTFET to the level of the nTFET. As both devices have similar SS, the VTC is again well balanced. More over, NMs look quite good. This suggests that steep SS results in worse NM. The reason is however in the slow onset of the output characteristics of the TFETs as shown for pTFET in Fig. 5.8a, which gets even slower for higher gate-source bias V_{GS} or V_{DD} (Fig. 5.10). This means that the typical way to make MOSFET-based SRAM work, V_{DD} increase, is not applicable for TFETs. This tricks works for regular MOSFETs as they exhibit the opposite behaviour: NMs improve with increase of V_{DD} [128]. Notice, however, that both MOSFETs and TFETs fail to work well with V_{DD} lower than approximately 150 mV.

The last method of balancing n- and p-TFET by modification of the pTFET work function seems to be the best (Fig. 5.9). The VTC is perfectly balanced thanks to the careful selection of the work-function value. The effective pTFET V_{GS} decreased due to this work function shift, which immediately resulted in the NM improvement. Therefore, in the remaining analysis, the pTFET work function is always adjusted to have balanced VTC. Doing this, we maximize the NMs and use realistic, rather then artificial devices. As pTFET has steeper slope than nTFET, its I_{OFF} always turns out to be smaller than that of nTFET after the work function adjustment. Another conclusion which we may draw out

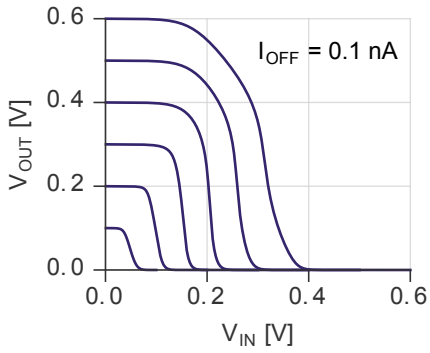


Figure 5.10: V_{DD} impact on VTC. The pTFET work function is adjusted such that the low NM is equal to high NM. Increase of V_{DD} results in the NM degradation.

of this study is that in general TFETs have to operate in a low leakage regime to minimize the negative impact of the delayed onset of the output characteristics.

5.5.4 Excessive energy consumption of a ring oscillator

So far, we looked at the static NMs of an inverter. Let us look now at the performance of TFET-based ring oscillator. We have used the same test bench as we used for the VFETs, that is to say 15-stages inverter-based ring oscillator with fan-out of three and 50CGP-long interconnect wires between the stages.

Figure 5.11a shows iso-frequency lines depending on V_{DD} and nTFET I_{OFF}. The pTFET I_{OFF} is always smaller, therefore the overall leakage is solely defined by the nTFET. Figure 5.11b shows iso-energy lines. This is more interesting than the iso-frequency lines, as clearly something happens in the high-performance corner (high I_{OFF} and high V_{DD}). From the CV curves of our TFETs, we may see that the choice of I_{OFF} should barely change the device capacitance. The parasitic capacitance are bias independent. Therefore, energy per switch should be roughly constant for a given V_{DD} regardless of I_{OFF} choice. This is the case everywhere (vertical iso-energy lines), except the high-performance corner, where energy lines start to bend. Essentially, this means that the ring oscillator consumes some excessive energy. Where may it come from?

During the switching, some short-circuit current always flows though the devices. However, typically, it flows for the very short time and might be neglected. This

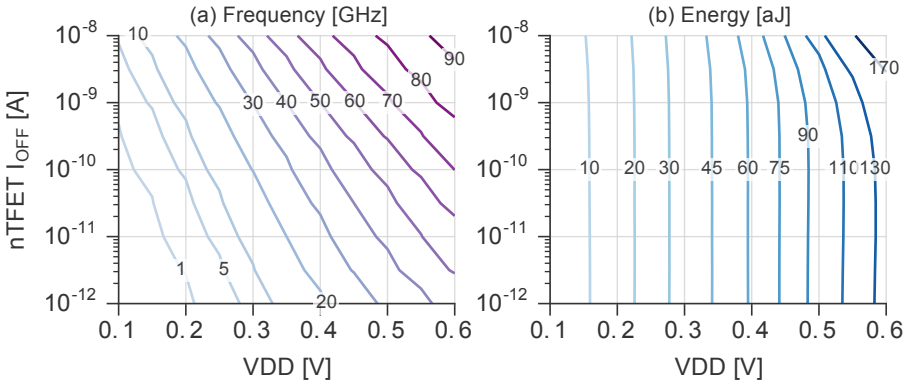


Figure 5.11: (a) TFET RO iso-frequency lines as a function of nTFET I_{OFF} and V_{DD}. (b) TFET RO iso-energy lines as a function nTFET I_{OFF} and V_{DD}. pTFET I_{OFF} is adjusted to have NML=NMH=NM and it is always lower than, the nTFET I_{OFF}. Energy lines start to bend in the high performance corner pointing that the excessive energy consumption happens somewhere.

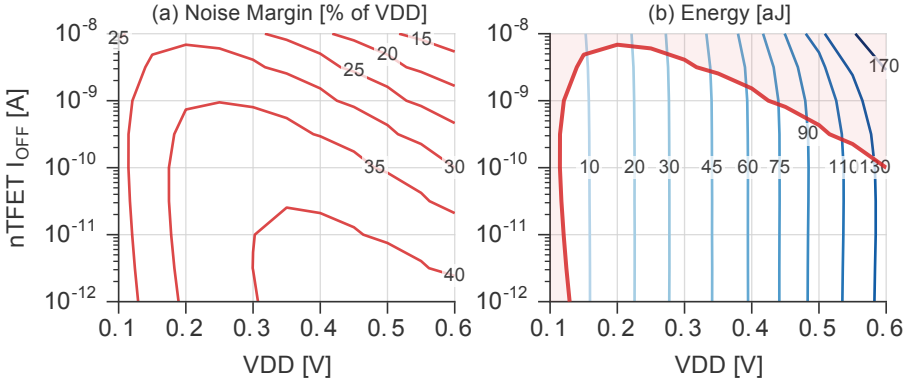


Figure 5.12: (a) TFET iso-NM lines as a function of nTFET I_{OFF} and V_{DD}. (b) TFET RO iso-energy lines as a function nTFET I_{OFF} and V_{DD}. Settling times are too long if NM is less than 30% (red line in (b)) of V_{DD}, resulting in excessive energy consumption. Region with too low NM (in red) defines the undesirable operating regime for TFETs.

is because inverter gain is relatively high and a transition from V_{DD} to V_{SS} and backwards happens abruptly. However, based on Fig. 5.10 we may see that this transition smears out with V_{DD} increase for TFETs. Gain is linked with NMs, therefore we plot iso-NM lines in Figure 5.12a as a function of V_{DD} and

Table 5.2: Impact of nTFET drain doping. $V_{D\ lin} = 50\,\text{mV}$, $V_{D\ sat} = 400\,\text{mV}$ and nTFET off-current is 10 pA.

N_{Drain} [cm^{-3}]	5E18	5E17
SS [mV/dec]	44.3	30.9
$I_{D\ lin}$ [μA]	2.50	1.51
$I_{D\ sat}$ [μA]	12.5	18.4
Frequency [GHz]	27.6	38.6
Energy [aJ]	61.7	61.7
NM [mV]	163	145

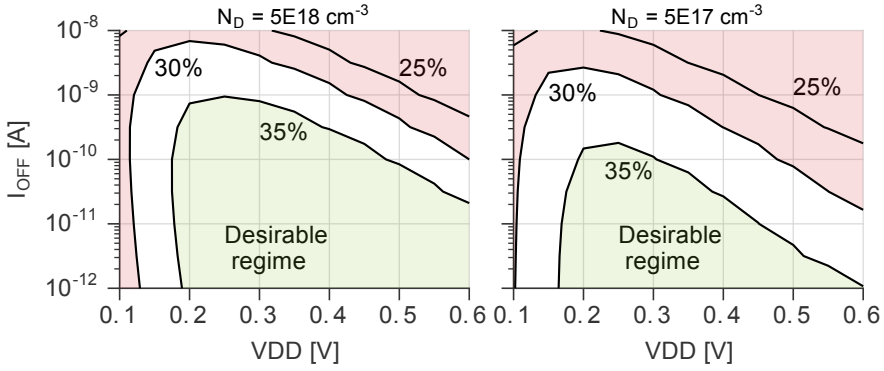


Figure 5.13: Impact of nTFET drain doping on noise margins depending on V_{DD} and I_{OFF}.

I_{OFF}. Values of NMs are given in percentage of V_{DD}, meaning that maximum possible value is 50%. Because of the pTFET work function adjustment, the low NM is equal to high NM, so we report a single a value. Notice that the link between NMs, I_{OFF} and V_{DD} is not that simple. However, if we put the plot with NM iso-lines on top of the plot with the energy iso-lines, we may directly see that energy iso-lines start to bend when NM values drop below 30%. This empirical constraint on NMs defines the $I_{OFF} - V_{DD}$ window, where our TFET devices may operate. Interestingly, the impact of poor VTCs is visible even for digital applications, which is rarely the case for regular MOSFETs.

We mentioned in section 5.5.1, that we had another nTFET device designed with lower drain doping ($5{\times}10^{17}\,\text{cm}^{-3}$). The impact of the lower drain doping is summarized in Table 5.2. Actually, it looks like the device with lower drain

doping is better as it brings almost 40% increase in frequency at the same energy per switch. This is true, as long as we operate in the safe zone, where NM is sufficiently high. However, for the lightly doped device this zone is significantly smaller as shown in Fig. 5.13. Therefore, we may see this as a trade-off between the control of the operating point and speed. Work function variations affect I_{OFF}, there are always some fluctuations on the value of the voltage supply, so it is more difficult to stay within the safe zone for lightly doped drain nTFET, although the operating frequency is higher.

5.6 Summary and conclusions

This chapter introduced the tunnel field-effect transistor (TFET) device as a possible option for the next generation low power switch. We described how we were able to fit the compact model against the quantum mechanical simulations to make circuit simulations possible. We identified a couple of issues related to TFETs operation in digital circuits: noise margins which get degraded with increase of V_{DD}, and non-negligible short-circuit currents if devices are pushed towards the high-performance regime (high leakage, and high V_{DD}). Both issues are related to the slow onset of TFET output characteristics. This is an important issue, which is often overlooked during the device design. We see that just playing with doping levels leads to a trade-off between transfer and output characteristics. Therefore, their co-optimization is a challenge which has to be addressed in the future work (*e.g.* with the broken-gap heterojunctions [35], [129], or double gate designs [130]). On top of that, it is essential to co-optimize n- and p-TFETs so that the voltage transfer characteristics (VTC) of TFET-based inverters comes out well balanced naturally, because the trick which we used for this study (pTFET work-function adjustment) is not applicable in real life, as the adjustments depend on the operating point (leakage target and V_{DD}), which may change dynamically. Nevertheless, with the devices which we have now, we may see that they behave good if run at moderate V_{DD} and low leakage. Therefore, in the next chapter we will benchmark them against regular MOSFETs.

The key conclusion from this chapter is that although the TFET community is typically focused on the reduction of the subthreshold slope, it is important to pay attention to the output characteristics: super-linear onset and high drain saturation voltage have negative impact on circuits performance.

CHAPTER 6

Benchmark of Vertical and Lateral Devices

6.1 How to conduct a fair benchmark?

6.1.1 Introduction

So far, we have only discussed vertical devices: either regular MOSFETs or TFETs. However, these days the standard industry device is a FinFET [1], [2], [24], [131]. It is a lateral device. Shift towards vertical architecture is quite drastic for both integration engineers and chip designers. Therefore, the industry should have a good motivation to implement such a shift. In case we demonstrate that there is a room for VFETs, these devices might be considered as a prominent option for the further CMOS scaling. Yet, this demonstration has to be objective and fair, which means that the benchmarking methodology should be carefully planned.

We introduced the test vehicle for technology benchmarking in section 4.3.2. This is a ring oscillator with certain interconnect load between the stages. In principle, given the fact that lateral and vertical devices are quite different, the interconnect load is not necessary the same for these two architectures. The standard cells are designed differently, the drive of these cells is different, which should result in different physical chip designs and, consequently, possibly different wire and/or fan-out distribution.

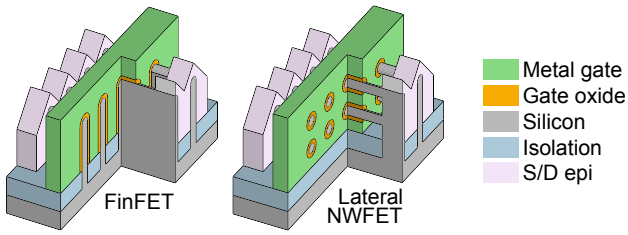


Figure 6.1: 3-D sketches of FinFET and lateral NWFET.

On top of that, the ground rules should not necessary be the same for both vertical and lateral devices. As discussed in the introduction to this thesis, typical technology scaling has involved simultaneous scaling of both CGP and MP. Yet, as the main problem of lateral devices scaling is linked to the CGP scaling, it might be easier to back up on it, but to scale MP more aggressively. It is somewhat different for the vertical devices. Therefore, if different metal pitches are used, not only the wiring distribution differs between the lateral and vertical architectures, but also the electrical properties of these wires.

Once all the assumptions are different for various technologies, the interpretation of benchmarking results becomes difficult. Therefore, in here, we try to conduct the analysis of different devices with as many assumptions being similar as possible. As such, we do not change metal pitch for vertical and lateral devices and assume the same wiring distribution, that said the interconnect load on the ring oscillator is equivalent to the load from 50 CGPs long wire, and fan-out is three for both architectures.

6.1.2 Selection of the best lateral device

To make the benchmarking fair, the vertical devices should be compared to the best possible lateral devices, as the industry would prefer to squeeze everything possible from the conventional lateral devices before it shifts towards disruptive vertical architecture. There are at least two devices which can be used for the lateral architecture at 7 nm node and beyond: FinFET, which is currently in mass production, and lateral NWFET (Fig. 6.1). The general discussion on which one is better is similar to the one we had for the vertical devices in chapter 4, when we compared NSh-like channels and NW-like channels. In the next sections of this chapter we will conduct a quantitative comparison of these two options across the two technological nodes as stated in Table 3.1. The best devices will be selected for a comparison with vertical devices.

The selection of the best lateral device is, however, not that simple. Both FinFET and NWFET may be and should be carefully optimized. We made such an attempt for the early benchmarking of vertical and lateral devices [71]. However, in here, we redo the exercise as certain critical assumptions have changed. First, we assume that the the cell height at the 7 nm node is not 9.0 tracks anymore but 7.5 tracks. Second, we demonstrated that the wrap-around contact allows to enhance the performance of lateral devices [19], [20], [81], [82]. We believe, this is one of the key boosters, which has to be considered for lateral devices scaling beyond the 7 nm node. Next, the benchmark itself is different. The new benchmark is a better representation of the real chip as it is based on the statistical data from the SoC design (see section 4.3.2). In addition, in the early benchmarking exercise we optimized devices solely for speed. However, energy consumption is equally, if not more, important. Finally, we used to re-optimize devices depending on the operating conditions: V_{DD} and off-current targets. Yet, this cannot happen in the real design: due to lithography constraints it would be easier and cheaper to have a single gate length for both high-performance and low-leakage flavours. Also, although the actual choice of nominal V_{DD} is hard to make at this stage, for sure it will be dynamically adjusted depending on the computation load. This is called dynamic voltage-frequency scaling (DVFS) — a widely used technique for the performance-energy co-optimization [132].

The device parameters subject to the optimization remain the same: spacer thickness, gate length, fin height or number of vertically stacked NWs. Before we proceed to the optimization, let us discuss some critical process assumptions. The doping level in the extensions of lateral devices is not as high as it is in vertical devices. Traditionally, the doping in the extensions has been defined by ion implantation. This approach does not work well for the tightly positioned high aspect-ratio fins. Therefore, other methods like dopants diffusion from PSG seem to be more appropriate, although they cannot provide very high doping levels [26]. We assume $5{\times}10^{19}\,\mathrm{cm}^{-3}$ to be a realistic value. Yet, the S/D regions underneath the contacts are epitaxially grown, which allows us to increase the doping level, either with the *in-situ* doping or with ion implantation. Two goals may be achieved by this doping boosting. One is an introduction of stress into the channel. Another one is access resistance reduction through a) lowering the semiconductor resistivity in the region under the S/D contacts and, more importantly, b) lowering specififc contact resistivity. We assume that the doping level in this part of lateral devices is the same as in vertical devices which results in similar specific contact resistivity ($5{\times}10^{-9}\,\Omega\,\mathrm{cm}^2$). Stress level is assumed to be 1.5 GPa which is consistent with the state-of-art TCAD studies [133].

Spacer thickness optimization for lateral devices differs from such an optimization

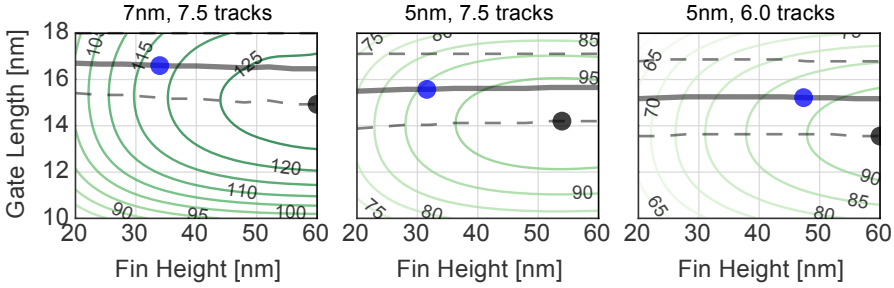


Figure 6.2: Iso-frequency lines as a function of fin height and gate length for different designs. Black dot indicate the maximum oscillation frequency. The blue dot indicate the gate length and fin height (and, thus, frequency) which we picked up for the further analysis. $I_{OFF} = 10\,\text{nA}$ and $V_{DD} = 0.6\,\text{V}$.

for vertical devices in two ways. First, thicker spacer means smaller contact size and/or gate length because of constraints coming from the fixed CGP. Second, as explained in the previous paragraph, resistivity of the extension is higher for lateral devices than for vertical devices. These two factors do not allow us to have thick spacers. At the same time, because of short spacers, parasitic capacitances are quite high and sensitive to spacer variations. A reasonable RC trade-off is achieved with 5 nm thick spacers. Spacer permittivity is 4.4, similar to the case of VFETs.

At first glance, the gate length optimization should happen simultaneously with fin height optimization. A taller fin results in higher current. Higher current causes larger IR-drop on access resistance. Shorter gate length allows to have larger room for S/D contacts, which lowers access resistance compensating the IR-drop. Let us see whether this simplified thinking is correct. Figure 6.2 shows iso-frequency contours depending on gate length and fin height for both 7 nm and 5 nm technological nodes at high-performance flavour ($I_{OFF} = 10\,\text{nA}$). Black dots indicate L_G - fin height combination corresponding to the highest frequency. Dashed lines passing through these dots indicate optimal gate lengths for each fin height. For the 7 nm node, our reasoning regarding L_G and fin height co-optimization is correct, although variations in optimal gate length are minor. For the 5 nm node, L_G - fin height line is completely flat. This may be explained by the fact that at the 5 nm node, scaled CGP pushes optimal gate length to smaller values than it was at the 7 nm node. As fin width remains the same, this push comes with the SS degradation. Thus, two processes compete when fin gets higher: smaller L_G for lower IR-drop and longer L_G for good SS.

Talking about SS, it is even more important for the low-leakage flavour ($I_{OFF} =$

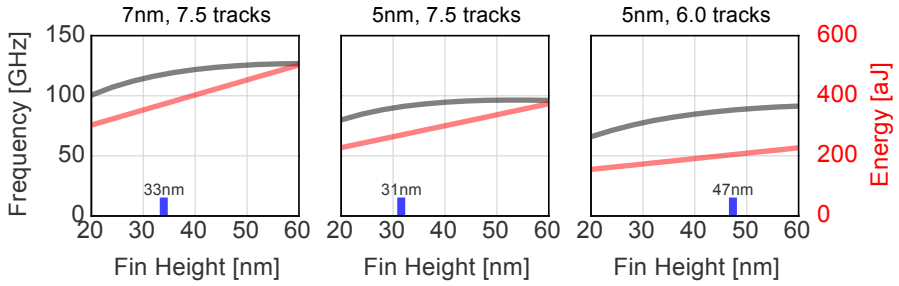


Figure 6.3: Frequency (black) and energy (red) as a function of fin height for different designs. Blue notches correspond to the fin heights which yield the lowest ED^{2}P. $I_{OFF} = 10\,\text{nA}$ and $V_{DD} = 0.6\,\text{V}$.

10 pA). The second dashed line on the top of each subplot in Fig. 6.2 indicated the gate which is optimal for this flavour. Notice here, that we did not sweep gate length beyond 18 nm. As we want to have a fixed gate length for both high and low V_{TH} flavours, we should find a compromise. The black thick solid line is just an average of two best L_G-s (at high and low V_{TH}). This line passes through several low V_{TH} iso-frequency lines. Which frequency (and, thus, fin height) to choose?

Figure 6.3 indicates how the ring oscillator (RO) frequency and energy per switch change with fin height for each device option. Frequency gets saturated with fin height increase, while energy per switch steadily increases. The reasons behind are similar to the ones explained in [71]: resistance, and, thus, drive, does not improve as fast as parasitic capacitance increase. This issue is somewhat less pronounced because of wrap-around contact as contacting surface increases together with fin height. Yet, there is non-negligible metal resistance between the fins, which gets higher with fin height increase [19].

It is quite common to co-optimize frequency and energy by minimizing energy-delay product (EDP) [134], [135]. However, in case the V_{DD} is varied (*e.g.* in the case of the DVFS), the optimization will not hold true. In that sense, minimization of ED^{2}P should yield better result as it is independent of the voltage in the first approximation [136]. This means, that the design may be optimized for the single metric. The voltage is then adjusted to obtain the required trade-off between delay and energy. Fin heights which correspond to the minimum of ED^{2}P are indicated in Fig. 6.3 with blue notches. Similarly, the blue circle is shown in Fig. 6.2.

Next, let us repeat the optimization exercise for the NWFET device. In a way, it is easier to optimize it because the fin height is quantized through a

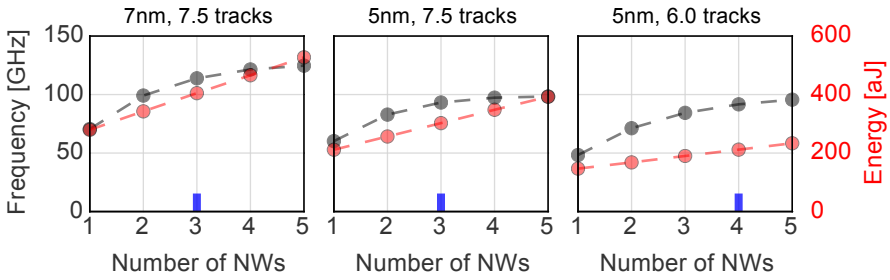


Figure 6.4: Frequency (black) and energy (red) as a function of number of stacked NWs for different designs. Blue notches correspond to the cases with the lowest ED^{2}P. $I_{OFF} = 10\,\text{nA}$ and $V_{DD} = 0.6\,\text{V}$.

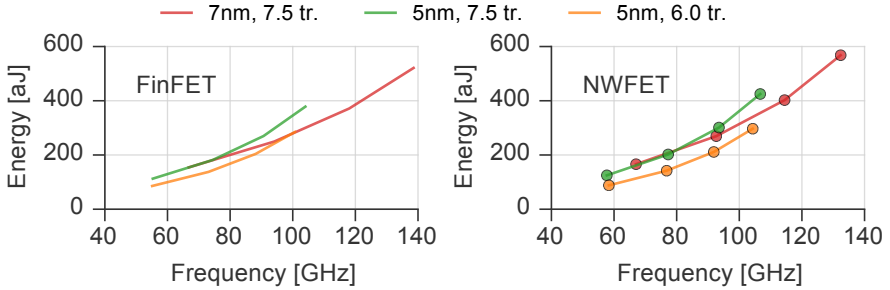


Figure 6.5: Energy and frequency characteristics of different a) FinFET-based designs, and b) NWFET-based designs. Supply voltage is varied from 0.4 V to 0.7 V, $I_{OFF} = 10\,\text{nA}$.

number of stacked NWs. Figure 6.4 is similar to the Fig. 6.3: for each number of stacked NWs the gate length is optimized between high and low V_{TH} flavours for maximum frequency which is reported in gray. The corresponding energy per switch is in red. Similarly to FinFETs, frequency saturates with an increase of number of stacked NWs. The optimal number of NWs is chosen based on the minimum ED^{2}P and is indicated with blue notches. This number is largely impacted by the vertical pitch between the NWs. We assumed it to be 14 nm. Tighter pitch would reduce parasitic capacitance, but the gate control would likely become worse because the whole gate stack will not fit in between the NWs.

Finally, Fig. 6.5 summarizes performance of selected FinFETs and NWFETs at various V_{DD} at both 7 nm and 5 nm nodes. At 5 nm node, energy per switch

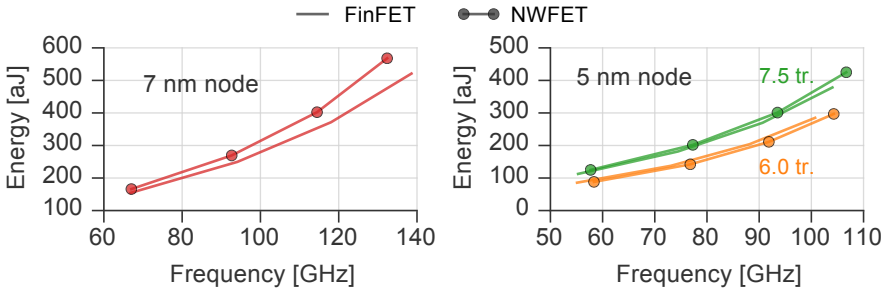


Figure 6.6: Energy and frequency characteristics of FinFETs and NWFETs made under a) 7 nm node ground rules, and b) 5 nm node ground rules. Supply voltage is varied from 0.4 V to 0.7 V, $I_{OFF} = 10$ nA.

improves with cell height scaling both for FinFETs and NWFETs, while the speed is maintained nearly the same. This means that the fin depopulation works quite efficiently. Two remarks regarding this observation. First, it should be possible to fabricate such high-aspect ratio fins (10:1). Second, the reader should keep in mind that we talk about the nominal device here. It is likely that due to variability, fin depopulation will not be as efficient as we see it in this study. Therefore, we keep both 7.5 tracks and 6.0 tracks options valid for the 5 nm node.

Fig. 6.6a compares performance of FinFETs and NWFETs at the 7 nm node, both made with 7.5 tracks tall standard cells. Introduction of NWFETs at this node is not necessary. Moreover, NWFETs consume more energy per switch at the high performance flavour than FinFETs. As for the low leakage flavour ($I_{OFF} = 10$ pA), NWFET' advantages over FinFET are minor. However, at 5 nm node (Fig. 6.6b), thanks to the better gate length scalability, a NWFET starts to outperform a FinFET. For the high-performance flavour, the 7.5 tracks design, there are still no benefits. However, for the 6.0 tracks design we see little benefits both in terms of speed at equal energy (+2.5%) and energy at equal speed (-6.6%). These benefits become more pronounced for the high V_{TH} flavour. Gain in speed at equal energy reaches 15.3%, and energy reduction at equal speed reaches 9.4%.

To conclude, for the 7 nm node, our reference point, we choose 7.5 tracks tall cells, FinFET-based design. However, for the 5 nm node we replace a device with a lateral NWFET and we assume that both 7.5 tracks and 6.0 tracks tall cells are possible. The final choice on the cell height depends on the variability parameters and tall fin process feasibility. Yet, even though a NWFET is somewhat better than a FinFET at 5 nm node, *it is still worse than*

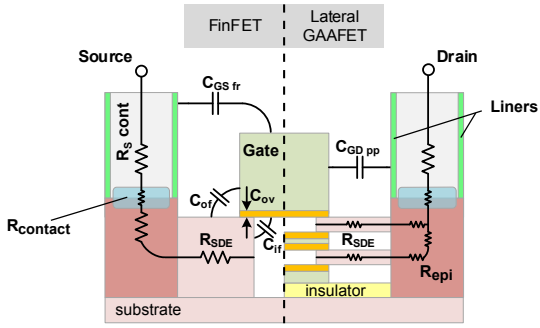


Figure 6.7: Sketches of source side of a FinFET and drain side of a lateral GAAFET devices showing main parasitics considered in the study.

the reference FinFET device from the 7 nm *node*. Next, we will see whether a vertical device can close this performance gap.

6.2 Parasitics comparison

Once lateral devices are chosen, we may proceed to the comparison with vertical devices, both regular MOSFETs and TFETs. First, let us compare the extrinsic *RC*-parasitics. To make our study easier to analyze, the BEOL load is kept the same across the devices (50CGP-long wires, with wire properties being dependent on the technological node).

There is a significant difference in device parasitics despite similar specific contact resistivity, spacer permittivity, *etc.* It originates from the fact that the architectures are different. The overview of main considered parasitics in lateral devices is shown in Fig. 6.7. Number of fins is dictated by layout. For the regular vertical MOSFETs, we use devices consisting of just three or two 7 nm thick channels (for 10-tracks and 8-tracks respectively). For the vertical TFETs we only look at 10-tracks cells with the slab length of 49 nm. In order to have a TFET device at 8-tracks tall cells, the slab should be 28 nm long. This is the regime where the double-gate approximation may be invalid, full 3D simulations are required to evaluate the performance of these devices (see section 5.2 on the limitation of the TFET quantum mechanical simulator).

The break-up of resistances for different devices is given in Fig. 6.8. Few things change when transition from the 7 nm node to the the 5 nm node. First, the fin channel is replaced by NWs, which results in higher extension resistance.

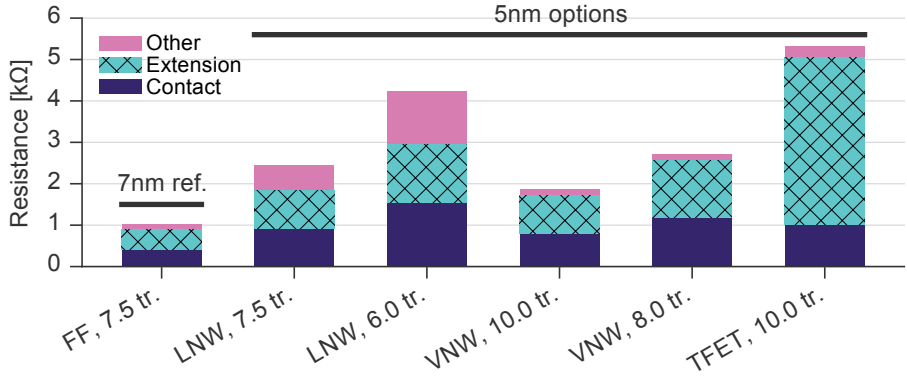


Figure 6.8: Parasitic resistance overview across various devices. Sum of source and drain resistances is indicated. "Other" includes spreading resistances, resistance in the epi underneath the contact, resistance of metal plugs, which get especially large for lateral devices made with 5 nm ground rules.

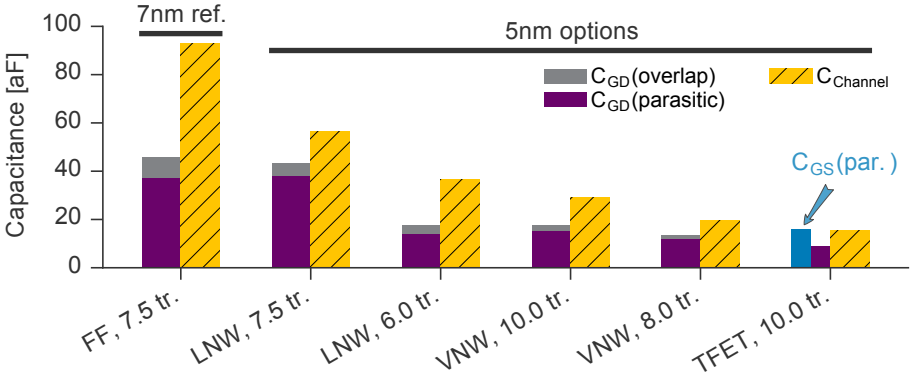


Figure 6.9: Capacitance overview across various devices. For all the devices except TFETs, $C_{GD} \approx C_{GS}$, therefore it is only C_{GD} which is reported. TFETs have long underlap on the drain side which lowers its capacitance. No overlap capacitance is reported for TFETs, as it is already included into the channel capacitance due to the nature of a simulator. All the values are the average between the values extracted for n- and p-channel devices.

This may be explained by the smaller cross-section of NWs in the region under the S/D spacer. Second, because of the reduced CGP, S/D contacts get very narrow, which results in high contact resistance. However, it is not only the contact resistance which gets higher, but also the metal plug resistance. These

plugs are typically fabricated of tungsten. This requires liners to be deposited first. They have a finite thickness which means that for the ultra narrow contact plugs there would be no tungsten core material left at all. This causes drastic increase in resistance. Vertical MOSFETs recover the situation by pushing access resistance back to the 7 nm level. TFETs still suffer from high resistance originating from their moderately doped drain extensions.

It is even more interesting to look at the capacitance difference (Fig. 6.9). Going from FinFETs to lateral NWFETs, and to vertical NWFETs, there is a steady decrease in channel capacitance, which is directly related to the decrease in the effective width. Same happens with an overlap capacitance, as it is directly proportional to the effective width. Interestingly, parasitic capacitance remain similar going from 7 nm FinFET to 5 nm NWFET in case the cell height remains the same (7.5 tracks) despite a fin depopulation. This may be explained by looking at the fin heights. For a FinFET, it is 33 nm. For a NWFET, it is $2 \times 14 + 7 + 10 = 45$ nm, where 14 nm is a vertical NW pitch, 7 nm is a NW diameter and 10 nm is an offset of the first NW from the substrate. As spacer thickness remains the same, for the first order calculations, we may assume the parasitic capacitance is proportional to a fin height and a number of fins. Thus, for a FinFET, $C \propto 3 \times 33 = 99$ nm, and for a NWFET, $C \propto 2 \times 45 = 90$ nm, which is quite close. Once we transition to the 6.0 tracks tall cells, parasitic capacitance drops by at least 1/3 ($C \propto 1 \times 59 = 59$ nm). In fact, the drop is even higher due to electrode extensions over the edges of a fin being smaller than half fin pitch. Yet, despite this drop, it is still hard to compete with vertical devices. This is because the VFET spacer thickness is largely relaxed with respect to lateral devices. TFETs have even smaller channel capacitance than any other devices, because they are made of III-V materials having relatively low density of states.

6.3 DC performance

As for the DC performance, we simply report the on-current for two flavours: high-performance ($I_{OFF} = 10$ nA) and low-leakage ($I_{OFF} = 10$ pA), and two V_{DD} values: 0.4 V and 0.6 V (Fig. 6.10 - 6.13). Whenever we operate in the high-performance regime, the role of subthreshold slope is quite limited, which makes lateral devices stronger than their competitors. This means that lateral devices will drive heavy load better than vertical devices. However, once we switch to the low-leakage flavour, TFETs demonstrate competitive currents at $V_{DD} = 0.6$ V. Moreover, they become way ahead of any conventional MOSFETs at $V_{DD} = 0.4$ V. Coupled with TFETs' low device capacitance, they should perform extremely good on the RO, as we will see it in the next sections.

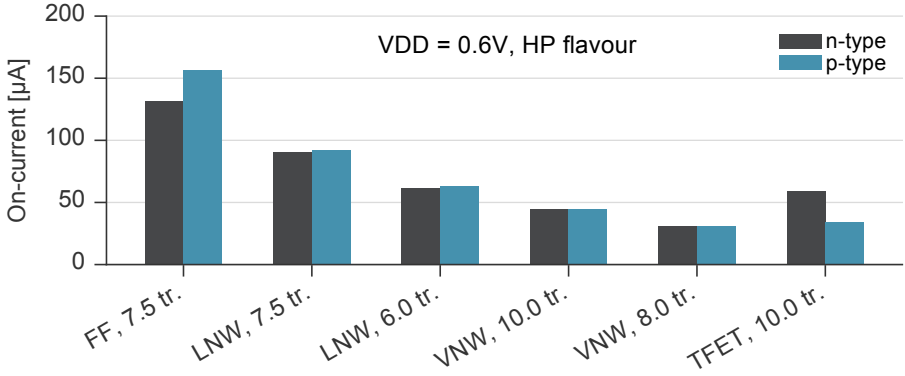


Figure 6.10: On-currents from different devices at $V_{DD} = 0.6\,\mathrm{V}$ made in high-performance flavour. Although, the TFET currents look quite decent, we know form the previous chapter, that the noise margins in this operating regime are unacceptably low.

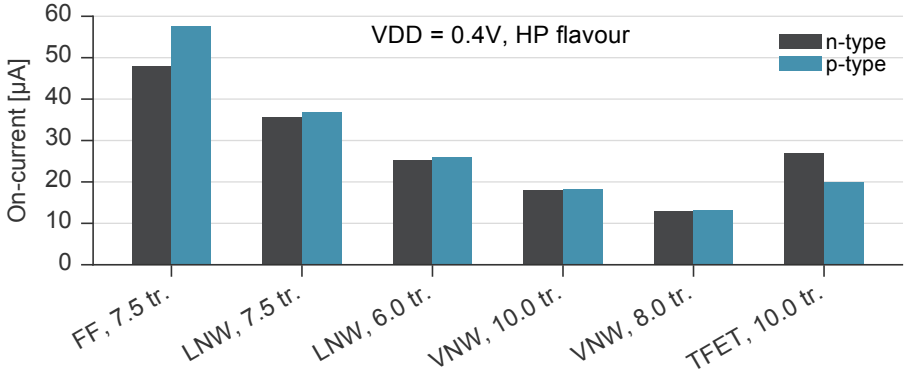


Figure 6.11: On-currents from different devices at $V_{DD} = 0.4\,\mathrm{V}$ made in high-performance flavour.

6.4 Performance on a ring oscillator

Given the four-fold higher TFET current in the low V_{DD}, low-leakage regime along with the lower device capacitance, it will not be a trivia task to plot TFET performance on the same plot with MOSFET performance. Therefore, we will split this section, discussing first the transition from lateral devices to vertical devices, and looking at TFETs right afterwards.

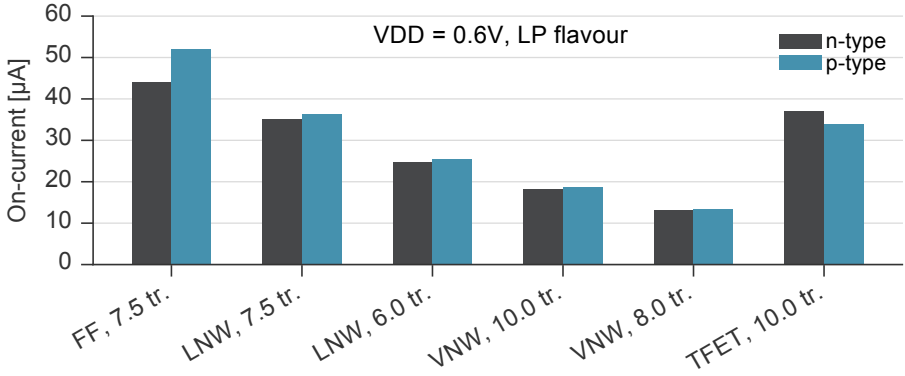


Figure 6.12: On-currents from different devices at $V_{DD} = 0.6\,\mathrm{V}$ made in low-leakage flavour.

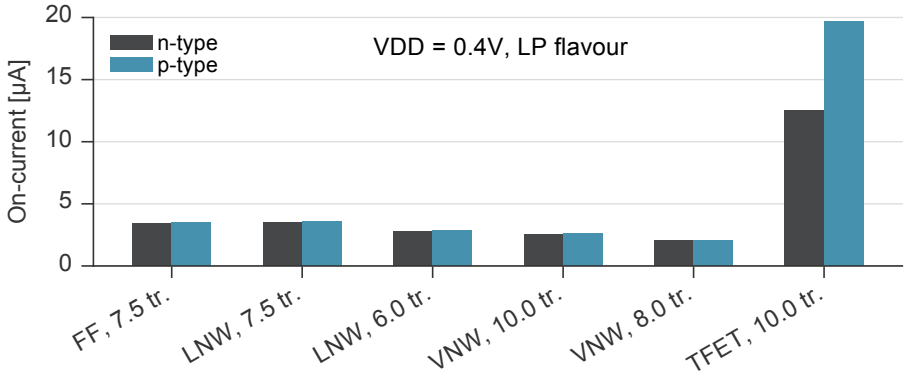


Figure 6.13: On-currents from different devices at $V_{DD} = 0.4\,\mathrm{V}$ made in low-leakage flavour. Thanks to the steep subthreshold slope, the TFET device is an absolute winner in this regime.

6.4.1 Lateral MOSFET against vertical MOSFET

In this section, we are going to combine the results obtained in the section 6.1.2 of this chapter together with the results for the vertical device from the chapter 4. On top of that we will look into the performance of 8.0 tracks tall vertical devices, which have only two NWs per device.

Figure 6.14 summarizes the results for the high-performance flavour. First of all, pure ground rules scaling of lateral devices brings power-performance *disadvantage*. However, if combined with fin depopulation, 5 nm node NWFET

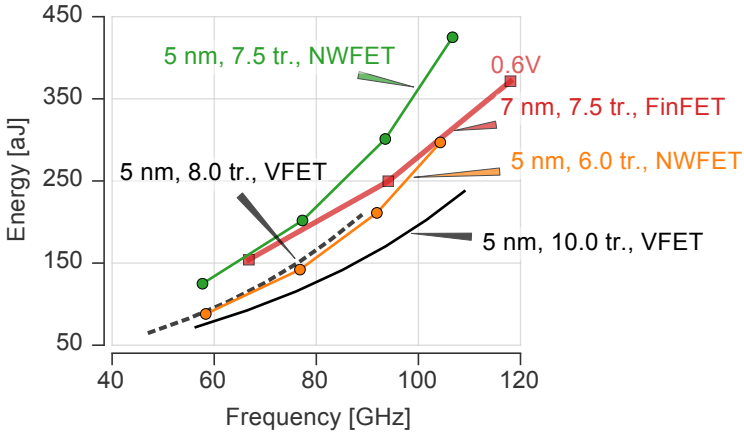


Figure 6.14: Benchmark of all the best 5 nm MOSFET candidatures with the reference 7 nm FinFET. Supply voltage is varied from 0.4 V to 0.7 V (to 0.6 V for the FinFET case), $I_{OFF} = 10$ nA.

may compete with FinFETs. Being more specific, at $V_{DD} = 0.7$ V, NWFET intersects FinFET's frequency-energy characteristics. for lower V_{DD} lateral NWFETs bring energy advantage at iso-performance. Interestingly, 7 nm node FinFET can reach equal performance at lower V_{DD}, but at higher energy per switch. As energy is proportional to $C \times V_{DD}^2$, we may conclude that capacitive load at 7 nm node is much higher than it is at 5 nm node. There are two main reasons for that. a) The FinFET device consists of three fins, while the NWFET consist of a single fin. b) The BEOL capacitive load is reduced at 5 nm node. The capacitance per unit length remains similar, but the wires themselves are shorter due to the scaling of the ground rules. Still, peak performance of the 7 nm node is not achievable with 5 nm node lateral NWFETs, or better say, NWFETs' energy consumption is going to be much higher than that of FinFETs, if NWFETs are pushed to the peak 7 nm frequencies .

Let us see what happens with VFETs. First, the 10 tracks tall cells. Their frequency-energy characteristics are parallel to those of the 7 nm node FinFET, with energy per switch being always smaller at iso-performance. At low V_{DD} energy savings are around 50%. Yet, at higher VDD they get less pronounced. As such, at $V_{DD} = 0.7$ V, VFETs consume 20% less energy. In order to reach speed of FinFETs, VFETs have to operate at higher V_{DD}.

The situation gets even worse for the 8 tracks tall cells. In this case, VFET devices consist of just two NWs each. The energy-frequency curve is shifted

mainly to the left, with high speed penalty and minor energy gains. The capacitive load did not improve as much as current degraded. This is a clear indication, that 8 tracks tall VFET is in the drive limited regime and it is not a feasible solution for the ultra-scaled 5 nm node unless some major innovation capable of drive enhancement happen.

The conclusion at this stage is pretty pessimistic. The lateral 7 nm node FinFET is very good. Conventional scaling toward 5 nm node through the CGP scaling does not work even in case the FinFET is replaced with the NWFET. The possible way out is a vertical device. Yet, we see that the advantages are not that striking and tend to be more pronounced for the modest performance applications. Moreover, vertical device does not seem well scalable, as it is already in the drive current limited regime. Nevertheless, as it is the best option for the 5 nm node so far, we pick it up for the benchmark against a TFET.

6.4.2 MOSFET against TFET

For the conventional MOSFETs benchmarking we mostly focused on the high-performance flavour. The reason behind is that we looked at the performance of different device options on a test-bench which represented an SoC critical path. The low-leakage flavour was covered by finding the gate length which would be a decent trade-off between the two flavours. As a result of this trade-off, the relative performance of the devices remained similar across the flavours (Fig. 4.21). This is different for a TFET. As its slope is much steeper and deviates largely from the MOSFET's 60 mV/dec, the relative performance between MOSFETs and TFETs will vary across the flavours. Once we start looking at different leakage floors, we have to modify our benchmark to take leakage into account in a fare way.

Probably the best way to introduce leakage into the picture is to play with activity factors. In case we target high-performance applications, like processors in data centers, leakage is not that important and activity factor is typically high. However, the situation is different for the mobile applications. Most of the time the device is the idle regime waiting for the computation. That said, the activity factor is quite low. Activity factor is a prefactor α for the active power consumption as stated in Eq. 6.1.

$$P_{total} = \alpha P_{active} + P_{leakage} \tag{6.1}$$

Thus, if α is low, the role of leakage power consumption increases with respect to the active power consumption. We look at two activity factors: $\alpha = 10\%$ and $\alpha = 0.1\%$. We let V_{TH} and V_{DD} to be freely optimized to reach minimum

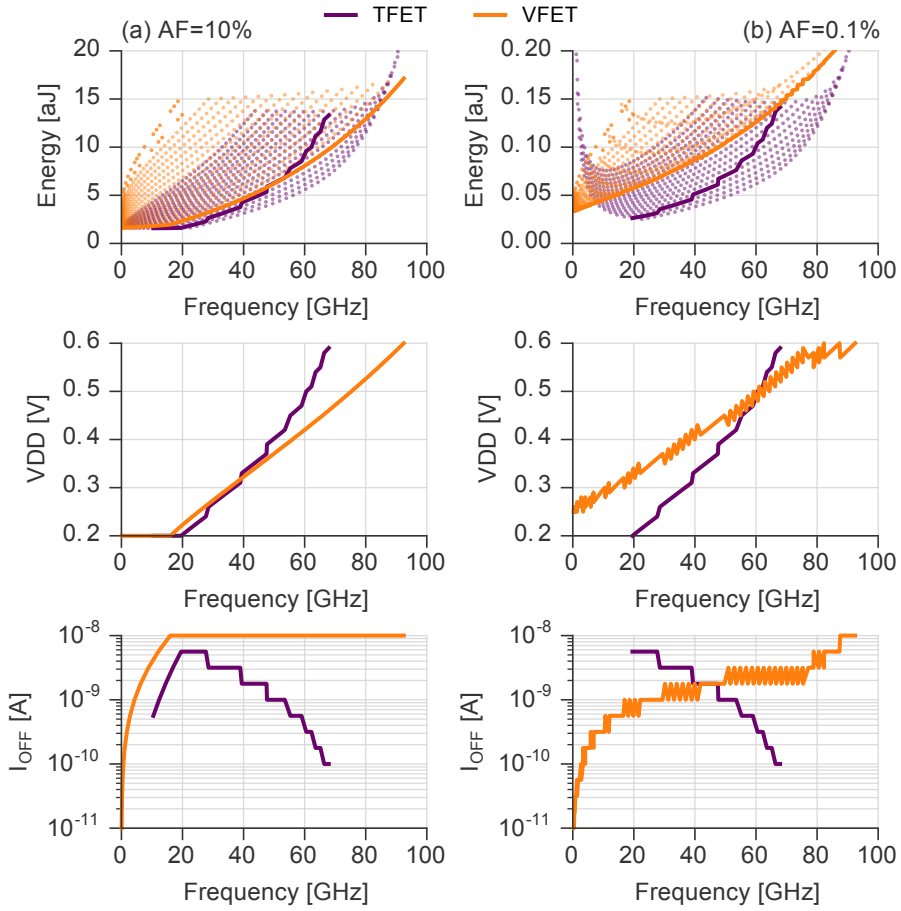


Figure 6.15: Energy-Frequency is plotted for a) 10% and b) 0.1% logic activity factor. V_{DD} range is 0.2 V to 0.6 V. I_{OFF} range is 10 pA to 10 nA. For each frequency, V_{DD} and I_{OFF} are chosen to minimize energy and reported in the corresponding subplots. Peak frequency for MOSFET is higher than for TFET.

energy consumption for a given frequency [137]. Optimum pair (V_{TH}, V_{DD}) depends on the activity factor. The two devices we benchmark are the TFET described in the previous chapter and the VFET consisting of three NWs per device.

Figure 6.15 summarizes the simulation results. First let us look at a conventional VFET device. Each point on the graph in the top row corresponds to a particular (V_{TH}, V_{DD}) pair. The thick solid line is a Pareto optimal curve: minimum

energy for each frequency. The two next rows indicate which values of I_{OFF} (linked with V_{TH}) and V_{DD} were chosen for a particular frequency. Same methodology was applied for the TFET device. However, some points from the solution cloud were filtered out based on the NM criterion (NM > 30% of V_{DD}). That is why the Pareto curve is not on the edge of the cloud, but within the cloud.

For the highly active applications ($\alpha = 10\%$), TFETs do not bring any feasible advantage over conventional MOSFETs even in case speed requirements are not stringent. Yet, in case the activity factor is relatively low ($\alpha = 0.1\%$), TFETs provide substantial energy gains. As such, for 20 GHz frequency, energy savings are more than 50%. Yet, there are few peculiar things here. First, the high performance regime is still not for TFETs unless NM drop to very small values. Second, 20 GHz frequency is *the best* frequency for TFETs at this testbench. Lower frequencies do not result in lower energy! This may be explained by leakage power taking over active power consumption. If we look at I_{OFF} trend versus frequency, we may notice that it steadily increases. We may interpret in the following way. As V_{DD} reduces, TFETs try to keep overdrive the same and lower their I_{OFF}, which is possible due to their steep SS. However, this strategy cannot be sustainable for ever. At certain moment the leakage power overcomes the active power. Partially, this was reinforced by the lower limit of V_{DD} sweep being 0.2 V.

Next, we decided to look at another TFET design. The one, where nTFET has lower drain doping ($5{\times}10^{17}\,\mathrm{cm}^{-3}$ instead of $5{\times}10^{18}\,\mathrm{cm}^{-3}$). We have already shown in the previous chapter that this option is much faster for the same I_{OFF} and V_{DD} condition, but suffers from smaller NMs. Figure 6.16 is similar to the Figure 6.15 but with different TFET device. The two TFETs demonstrate surprisingly similar behaviour with respect to their MOSFET competitor. A well visible difference between two TFETs is in the way these device react on I_{OFF} or V_{DD} variations. Looking at the solution cloud, the second option seems to be less sensitive to these variations. However this variation tolerance is apparent. We changed the I_{IOFF}, but not the V_{TH}. In the real life it is the V_{TH} which is going to be different. Therefore, yes, for the same I_{OFF} variations, the second options is better. However, for the same V_{TH} variations the second option will suffer from the larger I_{OFF} variations because of the steeper SS.

Let us have a close look at the difference between the two TFETs options (Fig. 6.17). In fact, both devices have very similar energy-speed characteristics. V_{DD} optimization is also similar. The difference is in the I_{OFF} choice. The option with nTFET having higher drain doping has a possibility to increase its leakage maintaining reasonable NMs (Fig. 5.13). Larger leakage helps to compensate for smaller drive.

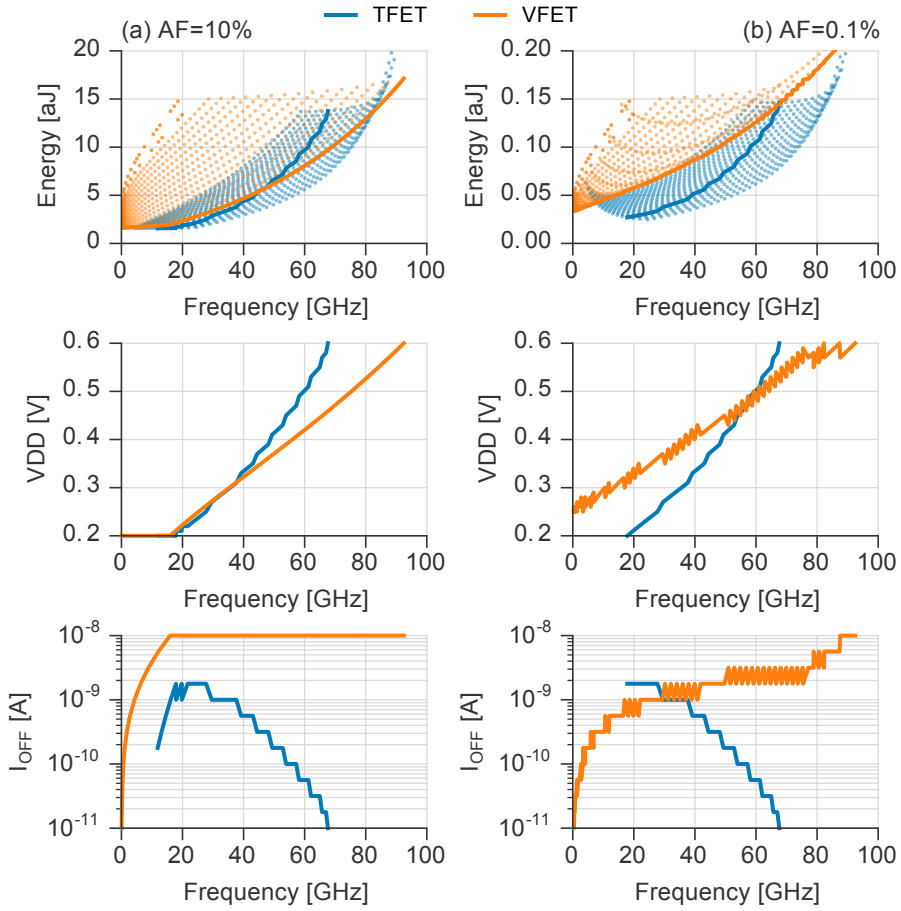


Figure 6.16: Energy-Frequency is plotted for a) 10% and b) 0.1% logic activity factor. V_{DD} range is 0.2 V to 0.6 V. I_{OFF} range is 10 pA to 10 nA. For each frequency, V_{DD} and I_{OFF} are chosen to minimize energy and reported in the corresponding subplots. Peak frequency for MOSFET is higher than for TFET. The difference with the Fig. 6.15 is in the used nTFET device. This nTFET has lower drain doping ($5{\times}10^{17}\,\mathrm{cm}^{-3}$ instead of $5{\times}10^{18}$).

At the end, we see that TFETs bring up to 50% energy savings with respect to MOSFETs if moderate speed and not-actively switching applications are targeted. In case NMs are less of an issue than what we assumed here, the application space for TFETs expands to the high-performance applications as well. However, very minor energy advantages are expected in this regime.

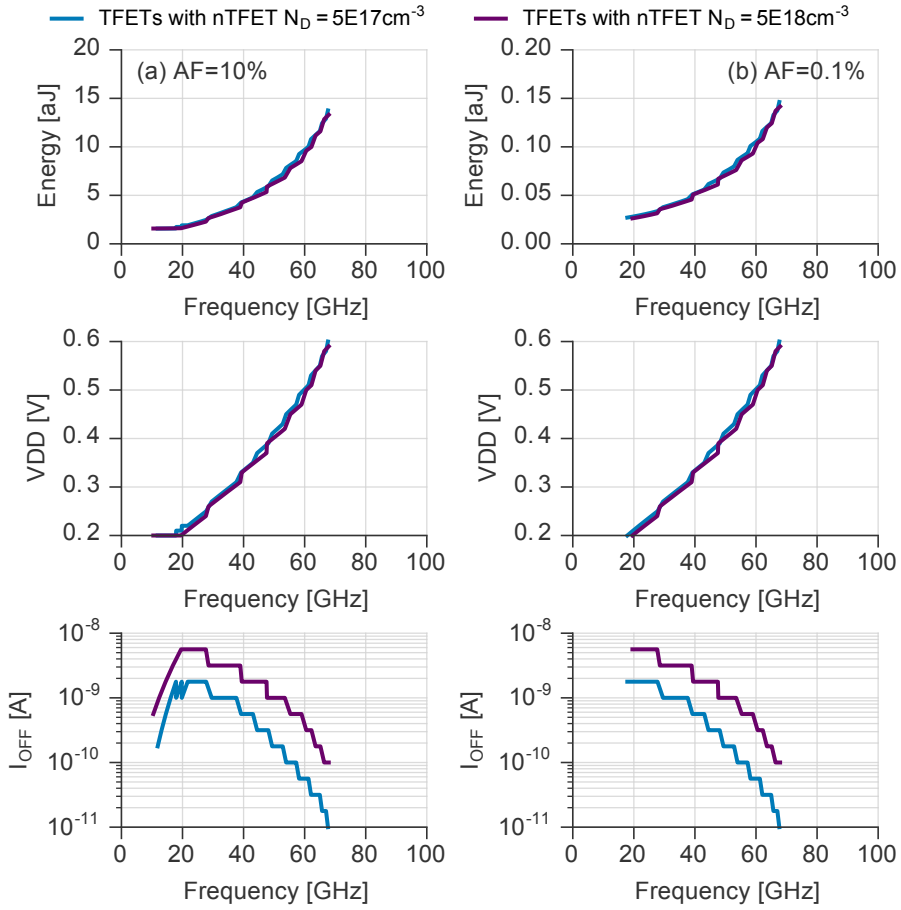


Figure 6.17: Performance comparison of two TFET options.

6.5 Summary and conclusions

This chapter suggested the way to make a fair benchmark of lateral and vertical architectures. We chose the FinFET device made with the 7 nm node design rules to be a reference point for our benchmarking. We could not identify any lateral device designed under the constraints of the 5 nm node which could compete with the 7 nm FinFET. Vertical MOSFETs seem to be able to enable further CMOS scaling down to the 5 nm node bringing almost 50% energy savings with respect to the 7 nm FinFET if run at moderate speeds. However, further scaling of vertical devices does not look any promising. Some advantages may come

if a regular vertical MOSFET is replaced by a vertical III-V heterojunction TFET. However, these advantages are only visible in the not actively switching applications and at relatively low switching speeds. TFETs devices may bring quite sensible benefits in case their noise margin do not get that much degraded with V_{DD} and I_{OFF} increase. Still the question of their scalability beyond the 5 nm node remains open.

The key take-away messages from this benchmark are the following.

- Lateral devices, FinFETs or NWFETs, are not scalable towards the the 5 nm node.
- Vertical devices, thanks to their smaller device parasitic capacitance bring power-performance advantage at the 5 nm node.
- Scaling beyond the 5 nm node requires vertical devices to have smaller active area, which results in the degraded current. Therefore, vertical architecture cannot provide sustainable scaling.
- Vertical TFETs, although bring some advantages over vertical MOSFETs at relatively low switching speeds, suffer from the same limited scalability as regular devices.

CHAPTER 7

Conclusions and Outlook

The Moore's law slowed down recently due to the issues with further CGP reduction and V_{DD} scaling. The lack of transistor scaling results in area penalties. Stagnating V_{DD} does not allow to efficiently reduce energy consumption. Performance can hardly be improved without smaller devices having lower energy per switch. The complexity of the nowadays technologies results in an increase of both R&D and manufacturing costs. All the four key metrics, Power-Performance-Area-Cost (PPAC), which have been improving thanks to the Moore's law are now under a pressure. Therefore, there is an urge need for the drastic change of a device.

In this work, we holistically evaluated the vertical architecture as a solution for further scaling. The hypothesis was that vertical devices had a limited dependency on the CGP and thus they were better suited for the scaling. Specifically, we focused on the 5 nm technological node, which should go into production around 2021. The ground rules were chosen to roughly follow the historical scaling trends: CGP = 32 nm and MP = 24 nm. To quantify the advantages which vertical devices might bring, we benchmarked them against lateral devices made with similar ground rules.

An inverter-based ring-oscillator used to be the main testbench for us. We set it up to represent a realistic high-performance mobile SoC by setting the fan-out of each stage to three and by adding the representative interconnect wire load between the stages. Although this a relative simple circuit, its simulation with TCAD tools would be very time consuming therefore we relied on the compact models. These models included not only the description of the carriers transport in the channel, but also the analytical description of *RC* parasitics

specific to a certain device. Device dimensions were based on actual layouts. The parasitics model was developed in a parametrized way and centered around these dimensions. As we looked into both conventional MOSFETs and TFETs, the intrinsic transport model had to be different. For MOSFETs, we based ourselves on the industry standard BSIM-CMG model [10] but extended it to support quasi-ballistic transport. For TFETs, we relied on the slightly extended model published in Ref. [109]. For both types of devices, we calibrated compact models to TCAD.

The results of our benchmarking exercise turned out to be rather pessimistic but interesting. Let us start with the non-electrical comparison, but just simple area comparison.

The often heard believe is that VFETs may bring vast improvements in area scaling. However this is true only in two cases. The first case is a simple inverter. VFETs based cells do not require dummy gates at the edges of the cells, while lateral devices use them to separate fins between adjacent cells. These dummy gates largely penalize an area of an inverter. In the case of more complex cells, dummy gates at their edges are relatively less important. Moreover, lateral devices start to win in area thanks to their ability to share source and drain contacts. This is not possible with vertical devices. Series connection between the vertical transistors is tricky to implement as it either requires flipping of S/D terminals for the adjacent devices (top to bottom, or the other way round), or deep vias which have to deliver signal from the bottom (*e.g.* drain of one transistor) to the top (*e.g.* source of another transistor). This implies the need for the extra routing tracks outside of active regions, which penalizes the area. The second case is an SRAM. Because of its high regularity, the wiring between the transistors may be designed in a smart way. Actually, if logic would be as regular as memory, it would be possible to make logic cells compact as well.

Still, if VFETs are compared with lateral devices at similar footprint, VFETs can bring improvements in power and performance, but these improvements are not as major as the associated change of architecture. Compared to the best lateral device at the 5 nm node (NWFET), the VFET is 13% faster at iso-energy per switch. Alternatively, it consumes 24% less energy per switch at iso-performance. The limited improvement in power and performance metrics is not the only factor which makes vertical devices somewhat repulsive. Introduction of VFETs is likely a one time scaling solution. To continue scaling beyond the 5 nm node, the CGP-MP product has to be scaled. The CGP reduction is limited by the channel thickness (already at its limit!) and the gate stack around it. Unless there is a solution to drastically scale the gate stack thickness, the CGP cannot be scaled much. Thus, all the area gains should come from the MP scaling and a consequent reduction of a standard cell height. This results in the degraded drive current. According to our analysis, scaling of cell height by 20% (10 tracks

to 8 tracks) results in 30% loss in device drive current and similar RO speed loss without major improvement in energy per switch as the RO gets dominated by middle of line (MOL) and BEOL parasitic capacitances.

This is, probably, the major finding of the work. Although the vertical architecture allows to break off the gate length and spacer thickness dependence on the CGP, the burden associated with the CGP scaling does not go away. Moreover, a VFET footprint scaling comes with a drive reduction unlike in lateral devices where a drive strength is modulated by a fin height or a number of stacked NWs, which are independent on a device footprint.

We made some other interesting observations during our exploration work. The performance improvement numbers reported a few lines above are valid for a typical mobile SoC. It would be fair to generalize, that mobile is all about power reduction. However, there are other applications with different device specifications. For example, high-performance computing servers or FPGA based logic. Both require high currents. In the first case, they are needed to push operating frequency, while in the second case, they are needed to drive long interconnect lines. Unfortunately, high current is not something what might be expected from the vertical devices. Effective width of these devices is limited by a footprint, so is a current. Traditional performance boosters, like S/D stressors, cannot be applied to a VFET because it is nearly impossible to obtain compressive or tensile strain. Therefore, the peak performance of the 5 nm node VFET is lower than that of the 7 nm node FinFET.

Nevertheless, in case of lightly loaded applications, which do not target too high frequencies, the VFETs stand out because of their small FEOL capacitance and not that severe impact of access resistance: contact size might be relaxed with respect to a lateral device, because it is also not constrained by the CGP. The VFET parasitic FEOL capacitance is relatively small because of the relaxed spacer thickness. This has a positive effect on both switching speed and power consumption.

Talking about power limited applications, it is logical to think first of the V_{DD} scaling. Conventional MOSFETs (no matter, vertical or lateral) are not well suited for V_{DD} scaling due to the fundamental limit on their subthreshold slope (SS) = 60mV/dec. V_{TH} cannot be scaled down without leakage increase. III-V heterojunction steep slope TFETs were evaluated as a possible step further from the VFETs, which would allow power reduction through V_{DD} scaling. Again, the results are not that impressive and very much application dependent.

In case the actively switching applications are targeted, the opportunity window where TFETs outperform MOSFETs is marginally small. However, in case of rarely switching applications there is quite some room for TFETs. We see that

TFETs consume less energy per switch at same speed if V_{DD} drops below 0.5 V. Energy savings may reach almost 50% for low frequencies. This phenomenon is easy to interpret. Low activity factors imply that leakage power is a significant fraction of the overall power consumption. TFETs allow to lower off-current maintaining similar performance. Interestingly, burst regimes are not well applicable to TFETs because at high V_{DD} their NMs severely degrade. The reason behind is related to the delayed onset of the output characteristics, which, in our analysis, is especially prominent for a p-type TFET.

Do we suggest that the vertical architecture has little to no future? The answer might be given only after the careful cost estimation associated with these devices along with the market analysis. These topics go beyond the scope of this work. We believe, that a closer look to VFETs is still needed, because electrically-wise there is an opportunity window for vertical devices, especially if we think of the upcoming power constrained Internet of things (IoT) era.

However, the reader has to keep in mind that we only evaluated the nominal device. What will happen with device performance once the variability is taken into account? Limited effective width will likely result in strong V_{TH} variations [138]. The nowadays fabrication process of vertical devices heavily relies on a number of deposition, planarization and etch-back cycles [139]. This results in difficulties to align positions of S/D junctions. Because of this drawback, it might be interesting to evaluate the PPAC metrics of vertical junction-less devices. Effect of junctions misalignment will have even more impact on TFETs given their complex device structure. Talking about TFETs, there should be some work done on improving the output characteristics of p-type device. One of the proposed options is a double-gate structure, but that makes the device design even more difficult and susceptible to variability.

Talking about TFETs, we indicated that their peak performance is still smaller than it is of MOSFETs. Partially, this is related to the relatively small currents of TFETs. In this work we boosted it by using source pockets, but there is another interesting idea of using the so-called radial TFETs [140]. Their design enhances tunnelling currents by increasing the physical area of tunnelling junction. The nice thing about this increase is that it happens in a vertical direction, thus it does not affect the device footprint much. For sure, this TFET device concept should be assessed in great details.

In addition, we conducted our benchmark with an assumption of the same BEOL load between lateral and vertical architectures. This is still to be proven through the place and route exercises. The ground rules were the same between the devices, while this is not necessarily the right choice. Given different optimization processes and different design constraints, the ground rules should be co-optimized independently for lateral devices, and for vertical devices. In

this case, the BEOL load will definitely be different between the architectures.

For a complete technology assessment, other effects like self-heating should be taken into account. As the NWs are position vertically at a tight pitch this might be a major issue. Actually, TFETs might be superior here thanks to their lower sensitivity to temperature [34]. The list of factors to account for before the final conclusion regarding the technology introduction might be extended even further: reliability, noise, *etc.* Assessment of these factors is still a future work. Yet another possible way of scaling is a shift towards true 3D logic. In this thesis, we only looked at single layer of vertical devices. What if we start stacking devices on top of each other? The problem of accessing bottom most layers will be even more pronounced, but it might well happen that the area gains from stacking will be higher than the interconnect-related penalties. Actually, for the SRAM, it was predicted already that the bit cell stacking allows area scaling [141].

Therefore, at this moment, we recommend to investigate vertical devices deeper, as they have some potential to be an option for the ultimate CMOS scaling.

Bibliography

[1] H.-J. Cho, H. Oh, K. Nam, Y. Kim, K. Yeo, W. Kim, Y. Chung, Y. Nam, S. Kim, W. Kwon, M. Kang, I. Kim, H. Fukutome, C. Jeong, H. Shin, Y. Kim, D. Kim, S. Park, H. Oh, J. Jeong, S. Kim, D. Ha, J. Park, H. Rhee, S. Hyun, D. Shin, D. Kim, H. Kim, S. Maeda, K. Lee, Y. Kim, M. Kim, Y. Koh, B. Yoon, K. Shin, N. Lee, S. Kangh, K. Hwang, J. Lee, J.-H. Ku, S. Nam, S. Jung, H. Kang, J. Yoon, and E. Jung, "Si FinFET based 10nm technology with multi Vt gate stack for low power and high performance applications", in *2016 IEEE Symposium on VLSI Technology*, IEEE, Jun. 2016, pp. 1–2 (cit. on pp. 2, 64, 89).

[2] S.-Y. Wu, C. Lin, M. Chiang, J. Liaw, J. Cheng, C. Chang, V. Chang, K. Pan, C. Tsai, C. Yao, T. Miyashita, Y. Wu, K. Ting, C. Hsieh, R. Tsui, R. Chen, C. Yang, H. Chang, C. Lee, K. Chen, Y. Ku, and S. M. Jang, "Demonstration of a sub-0.03 μm^2 high density 6-T SRAM with scaled bulk FinFETs for mobile SOC applications beyond 10nm node", in *2016 IEEE Symposium on VLSI Technology*, IEEE, Jun. 2016, pp. 1–2 (cit. on pp. 2, 89).

[3] G. E. Moore, "Cramming more components onto integrated circuits, Reprinted from Electronics, volume 38, number 8, April 19, 1965, pp.114 ff.", *IEEE Solid-State Circuits Newsletter*, vol. 20, no. 3, pp. 33–35, Sep. 2006 (cit. on p. 2).

[4] G. E. Moore, "Progress in digital integrated electronics", in *1975 International Electron Devices Meeting (IEDM)*, vol. 21, IEEE, 1975, pp. 11–13 (cit. on p. 2).

[5] M. Bohr, "10 nm Technology Announcement", Intel, Tech. Rep., 2016 (cit. on p. 3).

[6] A. Mallik, J. Ryckaert, A. Mercha, D. Verkest, K. Ronse, and A. V.-Y. Thean, "Maintaining Moore's law: enabling cost-friendly dimensional scaling", in *Extreme Ultraviolet (EUV) Lithography VI*, O. R. Wood and E. M. Panning, Eds., vol. 9422, Apr. 2015, 94221N (cit. on p. 3).

[7] J. Ryckaert, P. Raghavan, P. Schuddinck, H.-B. Trong, A. Mallik, S. S. Sakhare, B. Chava, Y. Sherazi, P. Leray, A. Mercha, J. Bömmels, G. R. McIntyre, K. G. Ronse, A. V.-Y. Thean, Z. Tökei, A. Steegen, and D. Verkest, "DTCO at N7 and beyond: patterning and electrical compromises and opportunities",

in *Design-Process-Technology Co-optimization for Manufacturability IX*, J. L. Sturtevant and L. Capodieci, Eds., vol. 9427, Mar. 2015, p. 94270C (cit. on p. 4).

[8] L. Liebmann, Jia Zeng, Xuelian Zhu, Lei Yuan, G. Bouche, and J. Kye, "Overcoming scaling barriers through design technology CoOptimization", in *2016 IEEE Symposium on VLSI Technology*, IEEE, Jun. 2016, pp. 1–2 (cit. on pp. 4, 39).

[9] K. J. Kuhn, "Considerations for Ultimate CMOS Scaling", *IEEE Transactions on Electron Devices*, vol. 59, no. 7, pp. 1813–1828, Jul. 2012 (cit. on pp. 5, 22).

[10] J.-P. Colinge, *FinFETs and Other Multi-Gate Transistors*. Boston, MA: Springer US, 2008, pp. 113–153 (cit. on pp. 5, 17, 110).

[11] V. Moroz, X. W. Lin, L. Smith, J. Huang, M. Choi, T. Ma, J. Liu, Y. Zhang, J. Kawa, and Y. Saad, "Power-performance-area engineering of 5nm nanowire library cells", *International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*, pp. 433–436, 2015 (cit. on p. 5).

[12] J. B. Chang, M. Guillorn, P. M. Solomon, C. Lin, S. U. Engelmann, A. Pyzyna, J. A. Ott, and W. E. Haensch, "Scaling of SOI FinFETs down to Fin Width of 4 nm for the 10nm technology node", in *2011 Symposium on VLSI Technology (VLSIT)*, Jun. 2011, pp. 12–13 (cit. on p. 5).

[13] V. Subramanian, A. Mercha, B. Parvais, J. Loo, C. Gustin, M. Dehan, N. Collaert, M. Jurczak, G. Groeseneken, W. Sansen, and S. Decoutere, "Impact of fin width on digital and analog performances of n-FinFETs", *Solid-State Electronics*, vol. 51, no. 4, pp. 551–559, Apr. 2007 (cit. on p. 5).

[14] C. Munkang, V. Moroz, L. Smith, and J. Huang, "Extending drift-diffusion paradigm into the era of FinFETs and nanowires", in *2015 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*, IEEE, Sep. 2015, pp. 242–245 (cit. on pp. 5, 56, 57).

[15] V. Moroz, S. Lee, H. Joanne, C. Munkang, T. Ma, L. Jie, Z. Yunqiang, L. Xi-Wei, J. Kawa, and Y. Saad, "Modeling and optimization of group IV and III-V FinFETs and nano-wires", in *2014 IEEE International Electron Devices Meeting*, IEEE, Dec. 2014, pp. 7.4.1–7.4.4 (cit. on pp. 5, 20).

[16] C. Auth and J. Plummer, "Scaling theory for cylindrical, fully-depleted, surrounding-gate MOSFET's", *IEEE Electron Device Letters*, vol. 18, no. 2, pp. 74–76, Feb. 1997 (cit. on p. 6).

[17] A. V.-Y. Thean, D. Yakimets, T. Huynh-Bao, P. Schuddinck, S. Sakhare, M. Garcia Bardon, A. Sibaja-Hernandez, I. Ciofi, G. Eneman, A. Veloso, J. Ryckaert, P. Raghavan, A. Mercha, A. Mocuta, Z. Tökei, D. Verkest, P. Wambacq, K. De Meyer, and N. Collaert, "Vertical device architecture for 5nm and beyond: Device & Circuit implications", in *2015 Symposium on VLSI Technology (VLSI Technology)*, vol. 2015-Augus, IEEE, Jun. 2015, T26–T27 (cit. on p. 6).

[18] S. Datta, R. Pandey, A. Agrawal, S. K. Gupta, and R. Arghavani, "Impact of Contact and Local Interconnect Scaling on Logic Performance", in *2014 Symposium on VLSI Technology (VLSIT)*, IEEE, Jun. 2014, pp. 140–141 (cit. on p. 6).

[19] D. Yakimets, D. Jang, P. Raghavan, G. Eneman, H. Mertens, P. Schuddinck, A. Mallik, M. Garcia Bardon, N. Collaert, A. Mercha, D. Verkest, A. V.-Y. Thean, and K. De Meyer, "Lateral NWFET optimization for beyond 7nm nodes", in *2015 International Conference on IC Design & Technology (ICICDT)*, IEEE, Jun. 2015, pp. 1–4 (cit. on pp. 6, 91, 93).

[20] P. Raghavan, M. Garcia Bardon, D. Jang, P. Schuddinck, D. Yakimets, J. Ryckaert, A. Mercha, N. Horiguchi, N. Collaert, A. Mocuta, D. Mocuta, Z. Tökei, D. Verkest, A. V.-Y. Thean, and A. Steegen, "Holistic device exploration for 7nm node", in *2015 IEEE Custom Integrated Circuits Conference (CICC)*, IEEE, Sep. 2015, pp. 1–5 (cit. on pp. 6, 39, 91).

[21] Jiseok Kim, P. J. Oldiges, Hui-feng Li, H. Niimi, M. Raymond, P. Zeitzoff, V. Kamineni, P. Adusumilli, Chengyu Niu, and F. Chafik, "Specific contact resistivity of n-type Si and Ge M-S and M-I-S contacts", in *2015 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*, IEEE, Sep. 2015, pp. 234–237 (cit. on pp. 6, 61).

[22] H. Yu, M. Schaekers, E. Rosseel, A. Peter, J.-G. Lee, W.-B. Song, S. Demuynck, T. Chiarella, J.-A. Ragnarsson, S. Kubicek, J. Everaert, N. Horiguchi, K. Barla, D. Kim, N. Collaert, A. V.-Y. Thean, and K. De Meyer, "1.5×10–9 Ωcm2 Contact resistivity on highly doped Si:P using Ge pre-amorphization and Ti silicidation", in *2015 IEEE International Electron Devices Meeting (IEDM)*, IEEE, Dec. 2015, pp. 21.7.1–21.7.4 (cit. on p. 7).

[23] H. Yu, M. Schaekers, A. Hikavyy, E. Rosseel, A. Peter, K. Hollar, F. A. Khaja, W. Aderhold, L. Date, A. J. Mayur, J.-G. Lee, K. M. Shin, B. Douhard, S. A. Chew, S. Demuynck, S. Kubicek, D. Kim, A. Mocuta, K. Barla, N. Horiguchi, N. Collaert, A. V.-Y. Thean, and K. De Meyer, "Ultralow-resistivity CMOS contact scheme with pre-contact amorphization plus Ti (germano-)silicidation", in *2016 IEEE Symposium on VLSI Technology*, IEEE, Jun. 2016, pp. 1–2 (cit. on p. 7).

[24] S. Natarajan, M. Agostinelli, S. Akbar, M. Bost, A. Bowonder, V. Chikarmane, S. Chouksey, A. Dasgupta, K. Fischer, Q. Fu, T. Ghani, M. Giles, S. Govindaraju, R. Grover, W. Han, D. Hanken, E. Haralson, M. Haran, M. Heckscher, R. Heussner, P. Jain, R. James, R. Jhaveri, I. Jin, H. Kam, E. Karl, C. Kenyon, M. Liu, Y. Luo, R. Mehandru, S. Morarka, L. Neiberg, P. Packan, A. Paliwal, C. Parker, P. Patel, R. Patel, C. Pelto, L. Pipes, P. Plekhanov, M. Prince, S. Rajamani, J. Sandford, B. Sell, S. Sivakumar, P. Smith, B. Song, K. Tone, T. Troeger, J. Wiedemer, M. Yang, and K. Zhang, "A 14nm Logic Technology Featuring 2nd-Generation FinFET Transistors, Air-Gapped Interconnects, Self-Aligned Double Patterning and a 0.0588 μm^2 SRAM cell size", in *2014 IEEE International Electron Devices Meeting*, IEEE, Dec. 2014, pp. 3.7.1–3.7.3 (cit. on pp. 7, 89).

[25] C.-H. Lin, R. Kambhampati, R. J. Miller, T. B. Hook, A. Bryant, W. Haensch, P. Oldiges, I. Lauer, T. Yamashita, V. Basker, T. Standaert, K. Rim, E. Leobandung, H. Bu, and M. Khare, "Channel doping impact on FinFETs for 22nm and beyond", in *2012 Symposium on VLSI Technology (VLSIT)*, IEEE, Jun. 2012, pp. 15–16 (cit. on p. 7).

[26] Y. Kikuchi, T. Chiarella, D. De Roest, T. Blanquart, A. De Keersgieter, K. Kenis, A. Peter, P. Ong, E. Besien, Z. Tao, M.-S. Kim, S. Kubicek, S. Chew, T. Schram, S. Demuynck, A. Mocuta, D. Mocuta, and N. Horiguchi, "Electrical Characteristics of P-type Bulk Si Fin Field-Effect Transistor Using Solid-Source Doping with 1-nm Phosphosilicate Glass", *IEEE Electron Device Letters*, vol. 37, no. 9, pp. 1–1, 2016 (cit. on pp. 7, 91).

[27] D. Benoit, J. Mazurier, B. Varadarajan, S. Chhun, S. Lagrasta, C. Gaumer, D. Galpin, C. Fenouillet-Beranger, D. Vo-Thanh, D. Barge, R. Duru, R. Beneyton, B. Gong, N. Sun, N. Chauvet, P. Ruault, D. Winandy, B. van Schravendijk, P. Meijer, and O. Hinsinger, "Interest of SiCO low k=4.5 spacer deposited at low temperature (400°C) in the perspective of 3D VLSI integration", in *2015 IEEE International Electron Devices Meeting (IEDM)*, IEEE, Dec. 2015, pp. 8.6.1–8.6.4 (cit. on p. 7).

[28] T. Ando, J. B. Chang, S. K. Kanakasabapathy, P. Kulkarni, T. E. Standaert, and T. Yamashita, *FinFET parasitic capacitance reduction using air gap*, 2014 (cit. on p. 7).

[29] H. Liu, D. K. Mohata, A. Nidhi, V. Saripalli, V. Narayanan, and S. Datta, "Exploration of vertical MOSFET and tunnel FET device architecture for Sub 10nm node applications", in *70th Device Research Conference*, IEEE, Jun. 2012, pp. 233–234 (cit. on pp. 7, 40).

[30] G. Larrieu and X.-L. Han, "Vertical nanowire array-based field effect transistors for ultimate scaling.", *Nanoscale*, vol. 5, no. 6, pp. 2437–41, Mar. 2013 (cit. on p. 7).

[31] S. Maheshwaram, S. K. Manhas, G. Kaushal, B. Anand, and N. Singh, "Vertical Nanowire CMOS Parasitic Modeling and its Performance Analysis", *IEEE Transactions on Electron Devices*, vol. 60, no. 9, pp. 2943–2950, Sep. 2013 (cit. on pp. 7, 22, 28).

[32] T. Huynh-Bao, D. Yakimets, J. Ryckaert, I. Ciofi, R. Baert, A. Veloso, J. Boemmels, N. Collaert, P. Roussel, S. Demuynck, P. Raghavan, A. Mercha, Z. Tökei, D. Verkest, A. V.-Y. Thean, and P. Wambacq, "Circuit and process co-design with vertical gate-all-around nanowire FET technology to extend CMOS scaling for 5nm and beyond technologies", in *2014 44th European Solid State Device Research Conference (ESSDERC)*, IEEE, Sep. 2014, pp. 102–105 (cit. on pp. 7, 40, 44).

[33] C. Pan, P. Raghavan, D. Yakimets, P. Debacker, F. Catthoor, N. Collaert, Z. Tökei, D. Verkest, A. V.-Y. Thean, and A. Naeemi, "Technology/System Codesign and Benchmarking for Lateral and Vertical GAA Nanowire FETs at 5-nm Technology Node", *IEEE Transactions on Electron Devices*, vol. 62, no. 10, pp. 3125–3132, Oct. 2015 (cit. on p. 7).

[34] A. M. Ionescu and H. Riel, "Tunnel field-effect transistors as energy-efficient electronic switches.", *Nature*, vol. 479, no. 7373, pp. 329–37, Nov. 2011 (cit. on pp. 8, 113).

[35] U. E. Avci, D. H. Morris, and I. A. Young, "Tunnel Field-Effect Transistors: Prospects and Challenges", *IEEE Journal of the Electron Devices Society*, vol. 3, no. 3, pp. 88–95, May 2015 (cit. on pp. 8, 74, 87).

[36] D. Cutaia, K. E. Moselund, H. Schmid, M. Borg, A. Olziersky, and H. Riel, "Complementary III–V heterojunction lateral NW Tunnel FET technology on Si", in *2016 IEEE Symposium on VLSI Technology*, IEEE, Jun. 2016, pp. 1–2 (cit. on p. 8).

[37] J. Hergenrother, G. Wilk, T. Nigam, F. Klemens, D. Monroe, P. Silverman, T. Sorsch, B. Busch, M. Green, M. Baker, M. Baker, T. Boone, M. Bude, N. Ciampa, E. Ferry, A. Fiory, S. Hillenius, D. Jacobson, R. Johnson, P. Kalavade, R. Keller, C. King, A. Kornblit, H. Krautter, J.-C. Lee, W. Mansfield, J. Miner, M. Morris, S.-H. Oh, J. Rosamilia, B. Sapjeta, K. Short, K. Steiner, D. Muller, P. Voyles, J. Grazul, E. Shero, M. Givens, C. Pomarede, M. Mazanec, and C. Werkhoven, "50 nm Vertical Replacement-Gate (VRG) nMOSFETs with ALD HfO_2 and Al_2O_3 Gate Dielectrics", in *International Electron Devices Meeting. Technical Digest (Cat. No.01CH37224)*, IEEE, 2001, pp. 3.1.1–3.1.4 (cit. on p. 8).

[38] D. Vasileska and S. M. Goodnick, "Computational Electronics", *Synthesis Lectures on Computational Electromagnetics*, vol. 1, no. 1, pp. 1–216, Jan. 2006 (cit. on p. 14).

[39] A. N. Bhoj, R. V. Joshi, and N. K. Jha, "3-D-TCAD-Based Parasitic Capacitance Extraction for Emerging Multigate Devices and Circuits", *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 21, no. 11, pp. 2094–2105, Nov. 2013 (cit. on p. 15).

[40] S. Liu, K. Hsu, and P. Subramaniam, "ADMIT-ADVICE modeling interface tool", in *Proceedings of the IEEE 1988 Custom Integrated Circuits Conference*, IEEE, 1988, pp. 6.6/1–6.6/4 (cit. on p. 16).

[41] B. Wan, B. Hu, L. Zhou, and C.-J. Shi, "MCAST: an abstract-syntax-tree based model compiler for circuit simulation", in *Proceedings of the IEEE 2003 Custom Integrated Circuits Conference, 2003.*, IEEE, 2003, pp. 249–252 (cit. on p. 16).

[42] *Compact Model Coalition* (cit. on p. 16).

[43] C. C. McAndrew, G. J. Coram, K. K. Gullapalli, J. R. Jones, L. W. Nagel, A. S. Roy, J. Roychowdhury, A. J. Scholten, G. D. J. Smit, X. Wang, and S. Yoshitomi, "Best Practices for Compact Modeling in Verilog-A", *IEEE Journal of the Electron Devices Society*, vol. 3, no. 5, pp. 383–396, Sep. 2015 (cit. on pp. 16, 76).

[44] M. Shur, "Low ballistic mobility in submicron HEMTs", *IEEE Electron Device Letters*, vol. 23, no. 9, pp. 511–513, Sep. 2002 (cit. on pp. 17, 20, 21).

[45] J. Bude, "MOSFET modeling into the ballistic regime", in *2000 International Conference on Simulation Semiconductor Processes and Devices (Cat. No.00TH8502)*, IEEE, 2000, pp. 23–26 (cit. on pp. 17, 56).

[46] C. Jungemann, T. Grasser, B. Neinhus, and B. Meinerzhagen, "Failure of Moments-Based Transport Models in Nanoscale Devices Near Equilibrium", *IEEE Transactions on Electron Devices*, vol. 52, no. 11, pp. 2404–2408, Nov. 2005 (cit. on p. 17).

[47] M. S. Lundstrom, *Fundamentals of carrier transport*, Second edition. Cambridge University Press, 2000 (cit. on p. 17).

[48] D. Caughey and R. Thomas, "Carrier mobilities in silicon empirically related to doping and field", *Proceedings of the IEEE*, vol. 55, no. 12, pp. 2192–2193, 1967 (cit. on p. 17).

[49] J. A. Cooper and D. F. Nelson, "High-field drift velocity of electrons at the Si–SiO2 interface as determined by a time-of-flight technique", *Journal of Applied Physics*, vol. 54, no. 3, p. 1445, 1983 (cit. on p. 17).

[50] Y. Taur, *Fundamentals of Modern VLSI Devices*, Second edition. 2009 (cit. on pp. 17, 18).

[51] A. Khakifirooz, O. M. Nayfeh, and D. Antoniadis, "A Simple Semiempirical Short-Channel MOSFET Current–Voltage Model Continuous Across All Regions of Operation and Employing Only Physical Parameters", *IEEE Transactions on Electron Devices*, vol. 56, no. 8, pp. 1674–1680, Aug. 2009 (cit. on p. 18).

[52] H. Pao and C. Sah, "Effects of diffusion current on characteristics of metal-oxide (insulator)-semiconductor transistors", *Solid-State Electronics*, vol. 9, no. 10, pp. 927–937, Oct. 1966 (cit. on p. 19).

[53] K. Natori, "Ballistic metal-oxide-semiconductor field effect transistor", *Journal of Applied Physics*, vol. 76, no. 8, p. 4879, 1994 (cit. on p. 19).

[54] Synopsys, *Sentaurus Device Monte Carlo, Sentaurus Band Structure User Guide, I-2013.12*, December. 2013 (cit. on pp. 19, 51).

[55] W. Guo, M. Choi, A. Rouhi, V. Moroz, G. Eneman, J. Mitard, L. Witters, G. Van der Plas, N. Collaert, G. Beyer, P. Absil, A. V.-Y. Thean, and E. Beyne, "Impact of 3D integration on 7nm high mobility channel devices operating in the ballistic regime", in *2014 IEEE International Electron Devices Meeting*, IEEE, Dec. 2014, pp. 7.1.1–7.1.4 (cit. on p. 20).

[56] L. Smith, C. Munkang, M. Frey, V. Moroz, A. Ziegler, and M. Luisier, "FinFET to nanowire transition at 5nm design rules", in *2015 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*, IEEE, Sep. 2015, pp. 254–257 (cit. on pp. 20, 53).

[57] K. Huet, J. Saint-Martin, A. Bournel, S. Galdin-Retailleau, P. Dollfus, G. Ghibaudo, and M. Mouis, "Monte Carlo study of apparent mobility reduction in nano-MOSFETs", in *2007 37th European Solid State Device Research Conference (ESSDERC)*, Sep. 2007, pp. 382–385 (cit. on pp. 20, 21, 56).

[58] G. Bidal, D. Fleury, G. Ghibaudo, F. Boeuf, and T. Skotnicki, "Guidelines for MOSFET device optimization accounting for L-dependent mobility degradation", in *2009 Silicon Nanoelectronics Workshop (SNW)*, Jun. 2009, pp. 25–26 (cit. on p. 20).

[59] M. Zilli, D. Esseni, P. Palestri, and L. Selmi, "On the Apparent Mobility in Nanometric n-MOSFETs", *IEEE Electron Device Letters*, vol. 28, no. 11, pp. 1036–1039, Nov. 2007 (cit. on p. 21).

[60] J. L. Huguenin, G. Bidal, S. Denorme, D. Fleury, N. Loubet, A. Pouydebasque, P. Perreau, F. Leverd, S. Barnola, R. Beneyton, B. Orlando, P. Gouraud, T. Salvetat, L. Clement, S. Monfray, G. Ghibaudo, F. Boeuf, and T. Skotnicki, "Gate-all-around technology: Taking advantage of ballistic transport?", *Solid-State Electronics*, vol. 54, no. 9, pp. 883–889, 2010 (cit. on p. 21).

[61] R. Granzner, V. Polyakov, F. Schwierz, M. Kittler, R. Luyken, W. Rösner, and M. Städele, "Simulation of nanoscale MOSFETs using modified drift-diffusion and hydrodynamic models and comparison with Monte Carlo results", *Microelectronic Engineering*, vol. 83, no. 2, pp. 241–246, Feb. 2006 (cit. on p. 21).

[62] A. Dixit, A. Kottantharayil, N. Collaert, M. Goodwin, M. Jurczak, and K. De Meyer, "Analysis of the Parasitic S/D Resistance in Multiple-Gate FETs", *IEEE Transactions on Electron Devices*, vol. 52, no. 6, pp. 1132–1140, Jun. 2005 (cit. on p. 21).

[63] Wen Wu and M. Chan, "Gate Resistance Modeling of Multifin MOS Devices", *IEEE Electron Device Letters*, vol. 27, no. 1, pp. 68–70, Jan. 2006 (cit. on p. 21).

[64] W. Wu and M. Chan, "Analysis of Geometry-Dependent Parasitics in Multifin Double-Gate FinFETs", *IEEE Transactions on Electron Devices*, vol. 54, no. 4, pp. 692–698, Apr. 2007 (cit. on pp. 21, 28).

[65] J. Zhuge, R. Wang, R. Huang, X. Zhang, and Y. Wang, "Investigation of Parasitic Effects and Design Optimization in Silicon Nanowire MOSFETs for RF Applications", *IEEE Transactions on Electron Devices*, vol. 55, no. 8, pp. 2142–2147, Aug. 2008 (cit. on p. 21).

[66] A. Scholten, G. Smit, R. Pijper, L. Tiemeijer, A. Mercha, and D. Klaassen, "FinFET compact modelling for analogue and RF applications", in *2010 International Electron Devices Meeting*, IEEE, Dec. 2010, pp. 8.4.1–8.4.4 (cit. on p. 21).

[67] J. T. Smith, Y. Zhao, C. Yang, and J. Appenzeller, "Effects of nanoscale contacts to silicon nanowires on contact resistance: Characterization and Modeling", in *68th Device Research Conference*, IEEE, Jun. 2010, pp. 139–140 (cit. on p. 21).

[68] N. Lu and R. A. Wachnik, "Modeling of Resistance in FinFET Local Interconnect", *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 62, no. 8, pp. 1899–1907, Aug. 2015 (cit. on p. 21).

[69] K. Jansson, E. Lind, and L.-E. Wernersson, "Performance Evaluation of III-V Nanowire Transistors", *IEEE Transactions on Electron Devices*, vol. 59, no. 9, pp. 2375–2382, Sep. 2012 (cit. on pp. 22, 41).

[70] S. Mohney, Y. Wang, M. Cabassi, K. Lew, S. Dey, J. Redwing, and T. Mayer, "Measuring the specific contact resistance of contacts to semiconductor nanowires", *Solid-State Electronics*, vol. 49, no. 2, pp. 227–232, Feb. 2005 (cit. on pp. 22, 23, 25).

[71] D. Yakimets, G. Eneman, P. Schuddinck, T. Huynh-Bao, M. Garcia Bardon, P. Raghavan, A. Veloso, N. Collaert, A. Mercha, D. Verkest, A. V.-Y. Thean, and K. De Meyer, "Vertical GAAFETs for the Ultimate CMOS Scaling", *IEEE Transactions on Electron Devices*, vol. 62, no. 5, pp. 1433–1439, May 2015 (cit. on pp. 23, 62, 91, 93).

[72] F. Léonard and a. A. Talin, "Electrical contacts to one- and two-dimensional nanomaterials.", *Nature nanotechnology*, vol. 6, no. 12, pp. 773–83, Dec. 2011 (cit. on p. 25).

[73] J. Lacord, G. Ghibaudo, and F. Boeuf, "Comprehensive and accurate parasitic capacitance models for two- and three-dimensional CMOS device structures", *IEEE Transactions on Electron Devices*, vol. 59, no. 5, pp. 1332–1344, 2012 (cit. on p. 28).

[74] Q. Xu, J. Zou, J. Luo, R. Wang, and R. Huang, "Predictive modeling of capacitance and resistance in gate-all-around cylindrical nanowire MOSFETs for parasitic design optimization", in *2010 10th IEEE International Conference on Solid-State and Integrated Circuit Technology*, IEEE, Nov. 2010, pp. 1958–1960 (cit. on p. 28).

[75] J. Zou, Q. Xu, J. Lou, R. Wang, R. Huang, and Y. Wang, "Predictive 3-D modeling of parasitic gate capacitance in gate-all-around cylindrical silicon nanowire MOSFETs", *IEEE Transactions on Electron Devices*, vol. 58, no. 10, pp. 3379–3387, Oct. 2011 (cit. on p. 28).

[76] E. Lind and L.-E. Wernersson, "Design of RF Properties for Vertical Nanowire MOSFETs", *IEEE Transactions on Nanotechnology*, vol. 10, no. 4, pp. 668–673, Jul. 2011 (cit. on p. 29).

[77] G. Hiblot, Q. Rafhay, F. Boeuf, and G. Ghibaudo, "Refined conformal mapping model for MOSFET parasitic capacitances based on elliptic integrals", *IEEE Transactions on Electron Devices*, vol. 62, no. 3, pp. 972–979, 2015 (cit. on p. 30).

[78] W. Hilberg, "From Approximations to Exact Relations for Characteristic Impedances", *IEEE Transactions on Microwave Theory and Techniques*, vol. 17, no. 5, pp. 259–265, May 1969 (cit. on p. 30).

[79] A. Jeffrey, *Advanced Engineering Mathematics*. Academic Press, 2002, p. 1160 (cit. on p. 30).

[80] M. Bohr, "14 nm Technology Announcement", Tech. Rep., 2014, pp. 14–38 (cit. on p. 39).

[81] S. C. Song, J. Xu, N. N. Mojumder, K. Rim, D. Yang, J. Bao, J. Zhu, J. Wang, M. Badaroglu, V. Machkaoutsan, P. Narayanasetti, B. Bucki, J. Fischer, and G. Yeap, "Holistic technology optimization and key enablers for 7nm mobile SoC", in *2015 Symposium on VLSI Circuits (VLSI Circuits)*, IEEE, Jun. 2015, T198–T199 (cit. on pp. 39, 91).

[82] M. Garcia Bardon, P. Schuddinck, P. Raghavan, D. Jang, D. Yakimets, A. Mercha, D. Verkest, and A. V.-Y. Thean, "Dimensioning for power and performance under 10nm: The limits of FinFETs scaling", in *2015 International Conference on IC Design & Technology (ICICDT)*, IEEE, Jun. 2015, pp. 1–4 (cit. on pp. 39, 91).

[83] L. T. Clark, V. Vashishtha, L. Shifren, A. Gujja, S. Sinha, B. Cline, C. Ramamurthy, and G. Yeric, "ASAP7: A 7-nm finFET predictive process design kit", *Microelectronics Journal*, vol. 53, pp. 105–115, Jul. 2016 (cit. on pp. 39, 44).

[84] B. Vandewalle, B. Chava, S. Sakhare, J. Ryckaert, and M. Dusa, "Design technology co-optimization for a robust 10nm Metal1 solution for logic design and SRAM", in *Proc. SPIE*, vol. 9053, Mar. 2014, 90530Q–90530Q–13 (cit. on p. 40).

[85] L.-A. Ragnarsson, H. Dekkers, P. Matagne, T. Schram, T. Conard, N. Horiguchi, and A. V.-Y. Thean, "Zero-thickness multi work function solutions for N7 bulk FinFETs", in *2016 IEEE Symposium on VLSI Technology*, IEEE, Jun. 2016, pp. 1–2 (cit. on pp. 40, 45).

[86] M. S. Kim, W. Cane-Wissing, X. Li, J. Sampson, S. Datta, S. K. Gupta, and V. Narayanan, "Comparative Area and Parasitics Analysis in FinFET and Heterojunction Vertical TFET Standard Cells", *ACM Journal on Emerging Technologies in Computing Systems*, vol. 12, no. 4, pp. 1–23, May 2016 (cit. on pp. 40, 41).

[87] D. Yakimets, T. H. Bao, M. Garcia Bardon, M. Dehan, N. Collaert, A. Mercha, Z. Tökei, A. V.-Y. Thean, D. Verkest, and K. De Meyer, "Lateral versus vertical gate-all-around FETs for beyond 7nm technologies", in *72nd Device Research Conference*, IEEE, Jun. 2014, pp. 133–134 (cit. on p. 41).

[88] T. Huynh-Bao, J. Ryckaert, S. Sakhare, A. Mercha, D. Verkest, A. V.-Y. Thean, and P. Wambacq, "Toward the 5nm technology: layout optimization and performance benchmark for logic/SRAMs using lateral and vertical GAA FETs", in *Proc. SPIE*, L. Capodieci and J. P. Cain, Eds., vol. 9781, Mar. 2016, p. 978 102 (cit. on p. 41).

[89] F. M. Bufler, F. O. Heinz, and L. Smith, "Efficient 3D Monte Carlo simulation of orientation and stress effects in FinFETs", in *2013 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*, IEEE, Sep. 2013, pp. 172–175 (cit. on p. 56).

[90] F. M. Bufler and L. Smith, "3D Monte Carlo simulation of FinFET and FDSOI devices with accurate quantum correction", *Journal of Computational Electronics*, vol. 12, no. 4, pp. 651–657, Dec. 2013 (cit. on p. 56).

[91] M. Lenzi, P. Palestri, E. Gnani, S. Reggiani, A. Gnudi, D. Esseni, L. Selmi, and G. Baccarani, "Investigation of the Transport Properties of Silicon Nanowires Using Deterministic and Monte Carlo Approaches to the Solution of the Boltzmann Transport Equation", *IEEE Transactions on Electron Devices*, vol. 55, no. 8, pp. 2086–2096, Aug. 2008 (cit. on p. 56).

[92] Synopsys, *Sentaurus Device User Guide, I-2013.12*. 2013 (cit. on p. 58).

[93] *International Technology Roadmap for Semiconductors*, 2013 (cit. on p. 61).

[94] G. Shine and K. C. Saraswat, "Limits of specific contact resistivity to Si, Ge and III-V semiconductors using interfacial layers", in *2013 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*, IEEE, Sep. 2013, pp. 69–72 (cit. on pp. 61, 79).

[95] C. Pan and A. Naeemi, "A Paradigm Shift in Local Interconnect Technology Design in the Era of Nanoscale Multigate and Gate-All-Around Devices", *IEEE Electron Device Letters*, vol. 36, no. 3, pp. 274–276, Mar. 2015 (cit. on p. 63).

[96] OpenCores, "http://opencores.org/", 2009 (cit. on pp. 63, 64).

[97] E. Kultursay, K. Swaminathan, V. Saripalli, V. Narayanan, M. T. Kandemir, and S. Datta, "Performance enhancement under power constraints using heterogeneous CMOS-TFET multicores", *Proceedings of the eighth IEEE/ACM/IFIP international conference on Hardware/software codesign and system synthesis - CODES+ISSS '12*, p. 245, 2012 (cit. on p. 64).

[98] S.-Y. Wu, C. Lin, M. Chiang, J. Liaw, J. Cheng, S. Yang, S. Chang, M. Liang, T. Miyashita, C. Tsai, C. Chang, V. Chang, Y. Wu, J. Chen, H. Chen, S. Chang, K. Pan, R. Tsui, C. Yao, K. Ting, T. Yamamoto, H. Huang, T. Lee, C. Lee, W. Chang, H. Lee, C. Chen, T. Chang, R. Chen, Y. Chiu, M. Tsai, S. M. Jang, K. Chen, and Y. Ku, "An enhanced 16nm CMOS technology featuring 2^{nd} Generation FinFET Transistors and Advanced Cu/low-k Interconnect for Low Power and High Performance Applications", in *2014 IEEE International Electron Devices Meeting*, IEEE, Dec. 2014, pp. 3.1.1–3.1.4 (cit. on p. 64).

[99] X. Jiang, S. Guo, R. Wang, X. Wang, B. Cheng, A. Asenov, and R. Huang, "A device-level characterization approach to quantify the impacts of different random variation sources in finfet technology", *IEEE Electron Device Letters*, vol. 37, no. 8, pp. 962–965, Aug. 2016 (cit. on p. 68).

[100] A. Verhulst, D. Verreck, Q. Smets, K.-H. Kao, M. Van de Put, R. Rooyackers, B. Sorée, A. Vandooren, K. De Meyer, G. Groeseneken, M. Heyns, A. Mocuta, N. Collaert, and A. V.-Y. Thean, "Perspective of tunnel-FET for future low-power technology nodes", in *2014 IEEE International Electron Devices Meeting*, IEEE, Dec. 2014, pp. 30.2.1–30.2.4 (cit. on pp. 73, 74).

[101] D. Verreck, A. S. Verhulst, K.-H. Kao, W. G. Vandenberghe, K. De Meyer, and G. Groeseneken, "Quantum Mechanical Performance Predictions of p-n-i-n Versus Pocketed Line Tunnel Field-Effect Transistors", *IEEE Transactions on Electron Devices*, vol. 60, no. 7, pp. 2128–2134, Jul. 2013 (cit. on p. 74).

[102] D. Verreck, A. S. Verhulst, B. Sorée, N. Collaert, A. Mocuta, A. V.-Y. Thean, and G. Groeseneken, "Improved source design for p-type tunnel field-effect transistors: Towards truly complementary logic", *Applied Physics Letters*, vol. 105, no. 24, p. 243 506, Dec. 2014 (cit. on p. 74).

[103] D. Verreck, A. S. Verhulst, M. Van de Put, B. Sorée, W. Magnus, A. Mocuta, N. Collaert, A. V.-Y. Thean, and G. Groeseneken, "Full-zone spectral envelope function formalism for the optimization of line and point tunnel field-effect transistors", *Journal of Applied Physics*, vol. 118, no. 13, p. 134 502, Oct. 2015 (cit. on p. 74).

[104] Q. Smets, D. Verreck, A. S. Verhulst, R. Rooyackers, C. Merckling, M. Van De Put, E. Simoen, W. Vandervorst, N. Collaert, A. V.-Y. Thean, B. Sorée, G. Groeseneken, and M. M. Heyns, "InGaAs tunnel diodes for the calibration of semi-classical and quantum mechanical band-to-band tunneling models", *Journal of Applied Physics*, vol. 115, no. 18, p. 184 503, May 2014 (cit. on p. 74).

[105] M. Garcia Bardon, H. P. Neves, R. Puers, and C. Van Hoof, "Pseudo-Two-Dimensional Model for Double-Gate Tunnel FETs Considering the Junctions Depletion Regions", *IEEE Transactions on Electron Devices*, vol. 57, no. 4, pp. 827–834, Apr. 2010 (cit. on p. 75).

[106] A. S. Verhulst, B. Sorée, D. Leonelli, W. G. Vandenberghe, and G. Groeseneken, "Modeling the single-gate, double-gate, and gate-all-around tunnel field-effect transistor", *Journal of Applied Physics*, vol. 107, no. 2, p. 024 518, 2010 (cit. on p. 75).

[107] A. Biswas, L. De Michielis, A. Bazigos, and A. M. Ionescu, "Compact modeling of DG-Tunnel FET for Verilog-A implementation", in *2015 45th European Solid State Device Research Conference (ESSDERC)*, IEEE, Sep. 2015, pp. 40–43 (cit. on p. 75).

[108] L. De Michielis, N. Dagtekin, A. Biswas, L. Lattanzio, L. Selmi, M. Luisier, H. Riel, and a. M. Ionescu, "An innovative band-to-band tunneling analytical model and implications in compact modeling of tunneling-based devices", *Applied Physics Letters*, vol. 103, no. 12, p. 123 509, 2013 (cit. on p. 75).

[109] H. Lu, D. Esseni, and A. Seabaugh, "Universal analytic model for tunnel FET circuit simulation", *Solid-State Electronics*, vol. 108, pp. 110–117, Jun. 2015 (cit. on pp. 75, 76, 110).

[110] T. Stensitzki, D. B. Allen, A. Ingargiola, and M. Newville, *LMFIT: Non-Linear Least-Square Minimization and Curve-Fitting for Python*, 2014 (cit. on p. 76).

[111] A. Parker, "Getting to the Heart of the Matter: Considerations for Large-Signal Modeling of Microwave Field-Effect Transistors", *IEEE Microwave Magazine*, vol. 16, no. 3, pp. 76–86, Apr. 2015 (cit. on p. 76).

[112] J. Giraldo, N. M. Vivas, E. Vila, and A. Badia, "Assessing the (a)symmetry of concentration-effect curves", *Pharmacology & Therapeutics*, vol. 95, no. 1, pp. 21–45, Jul. 2002 (cit. on p. 76).

[113] S. Mookerjea, R. Krishnan, S. Datta, and V. Narayanan, "On Enhanced Miller Capacitance Effect in Interband Tunnel Transistors", *IEEE Electron Device Letters*, vol. 30, no. 10, pp. 1102–1104, Oct. 2009 (cit. on p. 77).

[114] L. Zhang, J. He, and M. Chan, "A compact model for double-gate tunneling field-effect-transistors and its implications on circuit behaviors", in *2012 International Electron Devices Meeting*, IEEE, Dec. 2012, pp. 6.8.1–6.8.4 (cit. on p. 77).

[115] L. Zhang, X. Lin, J. He, and M. Chan, "An Analytical Charge Model for Double-Gate Tunnel FETs", *IEEE Transactions on Electron Devices*, vol. 59, no. 12, pp. 3217–3223, Dec. 2012 (cit. on p. 77).

[116] H. Liu, S. Datta, and V. Narayanan, "Steep switching tunnel FET: A promise to extend the energy efficient roadmap for post-CMOS digital and analog/RF applications", in *International Symposium on Low Power Electronics and Design (ISLPED)*, vol. 2, IEEE, Sep. 2013, pp. 145–150 (cit. on p. 77).

[117] N. Dagtekin and A. Ionescu, "Impact of Super-Linear Onset, Off-Region Due to Uni-Directional Conductance and Dominant CGD on Performance of TFET-Based Circuits", *IEEE Journal of the Electron Devices Society*, vol. 3, no. 3, pp. 233–239, May 2015 (cit. on pp. 77, 81).

[118] D. H. Morris, U. E. Avci, R. Rios, and I. A. Young, "Design of Low Voltage Tunneling-FET Logic Circuits Considering Asymmetric Conduction Characteristics", *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 4, no. 4, pp. 380–388, Dec. 2014 (cit. on p. 77).

[119] A. G. Lind, N. G. Rudawski, N. J. Vito, C. Hatem, M. C. Ridgway, R. Hengstebeck, B. R. Yates, and K. S. Jones, "Maximizing electrical activation of ion-implanted Si in $In_{0.53}Ga_{0.47}As$", *Applied Physics Letters*, vol. 103, no. 23, p. 232 102, 2013 (cit. on p. 79).

[120] A. K. Baraskar, M. a. Wistey, V. Jain, U. Singisetti, G. Burek, B. J. Thibeault, Y. J. Lee, A. C. Gossard, and M. J. W. Rodwell, "Ultralow resistance, nonalloyed Ohmic contacts to n-InGaAs", *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures*, vol. 27, no. 4, p. 2036, 2009 (cit. on p. 79).

[121] S. Sze and K. Ng, *Physics of Semiconductor Devices*, 3rd ed. Hoboken, NJ, USA: John Wiley & Sons, Inc., Oct. 2006 (cit. on p. 79).

[122] M. Berg, J. Svensson, E. Lind, and L.-E. Wernersson, "A transmission line method for evaluation of vertical InAs nanowire contacts", *Applied Physics Letters*, vol. 107, no. 23, p. 232 102, Dec. 2015 (cit. on p. 79).

[123] A. S. Verhulst, D. Leonelli, R. Rooyackers, and G. Groeseneken, "Drain voltage dependent analytical model of tunnel field-effect transistors", *Journal of Applied Physics*, vol. 110, no. 2, p. 024 510, 2011 (cit. on p. 80).

[124] B. Rajamohanan, D. Mohata, A. Ali, and S. Datta, "Insight into the output characteristics of III-V tunneling field effect transistors", *Applied Physics Letters*, vol. 102, no. 9, p. 092 105, 2013 (cit. on p. 80).

[125] L. De Michielis, L. Lattanzio, and A. M. Ionescu, "Understanding the Superlinear Onset of Tunnel-FET Output Characteristic", *IEEE Electron Device Letters*, vol. 33, no. 11, pp. 1523–1525, Nov. 2012 (cit. on p. 80).

[126] E. Gnani, S. Reggiani, A. Gnudi, and G. Baccarani, "Drain-conductance optimization in nanowire TFETs", in *2012 Proceedings of the European Solid-State Device Research Conference (ESSDERC)*, IEEE, Sep. 2012, pp. 105–108 (cit. on p. 80).

[127] A. Pal, A. B. Sachid, H. Gossner, and V. R. Rao, "Insights Into the Design and Optimization of Tunnel-FET Devices and Circuits", *IEEE Transactions on Electron Devices*, vol. 58, no. 4, pp. 1045–1053, Apr. 2011 (cit. on p. 81).

[128] R. Swanson and J. Meindl, "Ion-implanted complementary MOS transistors in low-voltage circuits", in *1972 IEEE International Solid-State Circuits Conference. Digest of Technical Papers*, IEEE, 1972, pp. 192–193 (cit. on p. 83).

[129] A. Afzalian, M. Passlack, and Y.-C. Yeo, "Atomistic Simulation of Gate-All-Around GaSb/InAs Nanowire TFETs Using a Fast Full-Band Mode-Space NEGF model", in *2016 International Symposium on VLSI Technology, Systems and Application (VLSI-TSA)*, IEEE, Apr. 2016, pp. 1–2 (cit. on p. 87).

[130] S. Saurabh and M. J. Kumar, "Novel Attributes of a Dual Material Gate Nanoscale Tunnel Field-Effect Transistor", *IEEE Transactions on Electron Devices*, vol. 58, no. 2, pp. 404–410, Feb. 2011. eprint: `1108.3148` (cit. on p. 87).

[131] D. Guo, G. Karve, G. Tsutsui, K.-Y. Lim, R. Robison, T. Hook, R. Vega, D. Liu, S. Bedell, S. Mochizuki, F. Lie, K. Akarvardar, M. Wang, R. Bao, S. Burns, V. Chan, K. Cheng, J. Demarest, J. Fronheiser, P. Hashemi, J. Kelly, J. Li, N. Loubet, P. Montanini, B. Sahu, M. Sankarapandian, S. Sieg, J. Sporre, J. Strane, R. Southwick, N. Tripathi, R. Venigalla, J. Wang, K. Watanabe, C. W. Yeung, D. Gupta, B. Doris, N. Felix, A. Jacob, H. Jagannathan, S. Kanakasabapathy, R. Mo, V. Narayanan, D. Sadana, P. Oldiges, J. Stathis, T. Yamashita, V. Paruchuri, M. Colburn, A. Knorr, R. Divakaruni, H. Bu, and M. Khare, "FINFET technology featuring high mobility SiGe channel for 10nm and beyond", in *2016 IEEE Symposium on VLSI Technology*, IEEE, Jun. 2016, pp. 1–2 (cit. on p. 89).

[132] T. Burd, T. Pering, A. Stratakos, and R. Brodersen, "A dynamic voltage scaled microprocessor system", *IEEE Journal of Solid-State Circuits*, vol. 35, no. 11, pp. 1571–1580, Nov. 2000 (cit. on p. 91).

[133] A. Gendron-Hansen, K. Korablev, I. Chakarov, J. Egley, J. Cho, and F. Benistant, "TCAD Analysis of FinFET Stress Engineering for CMOS Technology Scaling", in *2015 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*, IEEE, Sep. 2015, pp. 417–420 (cit. on p. 91).

[134] R. Gonzalez, B. Gordon, and M. Horowitz, "Supply and threshold voltage scaling for low power CMOS", *IEEE Journal of Solid-State Circuits*, vol. 32, no. 8, pp. 1210–1216, 1997 (cit. on p. 93).

[135] K. Nose and T. Sakurai, "Optimization of V_{DD} and V_{TH} for Low-Power and High Speed Applications", in *Proceedings of the 2000 conference on Asia South Pacific design automation - ASP-DAC '00*, New York, New York, USA: ACM Press, 2000, pp. 469–474 (cit. on p. 93).

[136] A. J. Martin, M. Nystrom, and P. Penzes, *Power Aware Computing*, R. Graybill and R. Melhem, Eds., ser. Series in Computer Science. Boston, MA: Springer US, 2002, pp. 293–299 (cit. on p. 93).

[137] M. Horowitz, E. Alon, D. Patil, S. Naffziger, R. Kumar, and K. Bernstein, "Scaling, power, and the future of CMOS", in *IEEE InternationalElectron Devices Meeting, 2005. IEDM Technical Digest.*, IEEE, 2005, pp. 9–15 (cit. on p. 103).

[138] M. Pelgrom, A. Duinmaijer, and A. Welbers, "Matching properties of MOS transistors", *IEEE Journal of Solid-State Circuits*, vol. 24, no. 5, pp. 1433–1439, Oct. 1989 (cit. on p. 112).

[139] A. Veloso, E. Altamirano-Sanchez, S. Brus, B. T. Chan, M. Cupak, M. Dehan, C. Delvaux, K. Devriendt, G. Eneman, M. Ercken, T. Huynh-Bao, T. Ivanov, P. Matagne, C. Merckling, V. Paraschiv, S. Ramesh, E. Rosseel, L. Rynders, A. Sibaja-Hernandez, S. Suhard, Z. Tao, E. Vecchio, N. Waldron, D. Yakimets, K. De Meyer, D. Mocuta, N. Collaert, and A. V.-Y. Thean, "Vertical nanowire fet integration and device aspects", *ECS Transactions*, no. 4, pp. 31–42, (cit. on p. 112).

[140] A. W. Dey, J. Svensson, M. Ek, E. Lind, C. Thelander, and L.-E. Wernersson, "Combining axial and radial nanowire heterostructures: Radial esaki diodes and tunnel field-effect transistors", *Nano Letters*, vol. 13, no. 12, pp. 5919–5924, 2013 (cit. on p. 112).

[141] A. Veloso, B. Parvais, P. Matagne, E. Simoen, T. Huynh-Bao, V. Paraschiv, E. Vecchio, K. Devriendt, E. Rosseel, M. Ercken, B. T. Chan, C. Delvaux, E. Altamirano-Sánchez, J. J. Versluijs, Z. Tao, S. Suhard, S. Brus, A. Sibaja-Hernandez, N. Waldron, P. Lagrain, O. Richard, H. Bender, A. Chasin, B. Kaczer, T. Ivanov, S. Ramesh, K. D. Meyer, J. Ryckaert, N. Collaert, and A. Thean, "Junctionless gate-all-around lateral and vertical nanowire fets with simplified processing for advanced logic and analog/rf applications and scaled sram cells", in *2016 IEEE Symposium on VLSI Technology*, Jun. 2016 (cit. on p. 113).

List of Figures

List of Tables

List of Publications

Journal papers

1. D. Yakimets, G. Eneman, P. Schuddinck, T. Huynh-Bao, M. Garcia Bardon, P. Raghavan, A. Veloso, N. Collaert, A. Mercha, D. Verkest, A. V.-Y. Thean, and K. De Meyer, "Vertical GAAFETs for the Ultimate CMOS Scaling", *IEEE Transactions on Electron Devices*, vol. 62, no. 5, pp. 1433–1439, May 2015.

2. T. Agarwal, D. Yakimets, P. Raghavan, I. Radu, A. V.-Y. Thean, M. Heyns, and W. Dehaene, "Benchmarking of MoS_2 FETs With Multigate Si-FET Options for 5 nm and Beyond", *IEEE Transactions on Electron Devices*, vol. 62, no. 12, pp. 4051–4056, Dec. 2015.

3. C. Pan, P. Raghavan, D. Yakimets, P. Debacker, F. Catthoor, N. Collaert, Z. Tökei, D. Verkest, A. V.-Y. Thean, and A. Naeemi, "Technology/System Codesign and Benchmarking for Lateral and Vertical GAA Nanowire FETs at 5-nm Technology Node", *IEEE Transactions on Electron Devices*, vol. 62, no. 10, pp. 3125–3132, Oct. 2015.

4. T. Huynh-Bao, S. Sakhare, D. Yakimets, J. Ryckaert, A. V.-Y. Thean, A. Mercha, D. Verkest, and P. Wambacq, "A Comprehensive Benchmark and Optimization of 5-nm Lateral and Vertical GAA 6T-SRAMs", *IEEE Transactions on Electron Devices*, vol. 63, no. 2, pp. 643–651, Feb. 2016.

5. M. Garcia Bardon, V. Moroz, G. Eneman, P. Schuddinck, M. Dehan, D. Yakimets, D. Jang, G. Van Der Plas, A. Mercha, A. V.-Y. Thean, D. Verkest, and A. Steegen, "Layout-induced stress effects in 14nm & 10nm FinFETs", *2013 Symposium on VLSI Technology (VLSIT)*, pp. 114–115, 2013.

6. A. V.-Y. Thean, N. Collaert, I. Radu, N. Waldron, C. Merckling, L. Witters, R. Loo, J. Mitard, R. Rooyackers, A. Vandooren, A. Verhulst, A. Veloso, D. Yakimets, T. Huynh-Bao, D. Chiappe, A. Vaysset, O. Zografos, M. Caymax, C. Huyghebaert, K. Barla, and A. Steegen, "Heterogeneous Nano- to Wide-Scale Co-Integration of Beyond-Si and Si CMOS Devices to Enhance Future Electronics", *ECS Transactions*, vol. 66, no. 4, pp. 3–14, May 2015.

7. A. Veloso, E. Altamirano-Sanchez, S. Brus, B. T. Chan, M. Cupak, M. Dehan, C. Delvaux, K. Devriendt, G. Eneman, M. Ercken, T. Huynh-Bao, T. Ivanov, P. Matagne, C. Merckling, V. Paraschiv, S. Ramesh, E. Rosseel, L. Rynders, A. Sibaja-Hernandez, S. Suhard, Z. Tao, E. Vecchio, N. Waldron, D. Yakimets, K. De Meyer, D. Mocuta, N. Collaert, and A. V.-Y. Thean, "Vertical nanowire fet integration and device aspects", *ECS Transactions*, no. 4, pp. 31–42,

Conference proceedings

1. D. Yakimets, T. Huynh-Bao, M. Garcia Bardon, M. Dehan, N. Collaert, A. Mercha, Z. Tökei, A. V.-Y. Thean, D. Verkest, and K. De Meyer, "Lateral versus vertical gate-all-around FETs for beyond 7nm technologies", in *72nd Device Research Conference*, IEEE, Jun. 2014, pp. 133–134.

2. D. Yakimets, D. Jang, P. Raghavan, G. Eneman, H. Mertens, P. Schuddinck, A. Mallik, M. Garcia Bardon, N. Collaert, A. Mercha, D. Verkest, A. V.-Y. Thean, and K. De Meyer, "Lateral NWFET optimization for beyond 7nm nodes", in *2015 International Conference on IC Design & Technology (ICICDT)*, IEEE, Jun. 2015, pp. 1–4.

3. T. Huynh-Bao, D. Yakimets, J. Ryckaert, I. Ciofi, R. Baert, A. Veloso, J. Boemmels, N. Collaert, P. Roussel, S. Demuynck, P. Raghavan, A. Mercha, Z. Tökei, D. Verkest, A. V.-Y. Thean, and P. Wambacq, "Circuit and process co-design with vertical gate-all-around nanowire FET technology to extend CMOS scaling for 5nm and beyond technologies", in *2014 44th European Solid State Device Research Conference (ESSDERC)*, IEEE, Sep. 2014, pp. 102–105.

4. A. V.-Y. Thean, D. Yakimets, T. Huynh-Bao, P. Schuddinck, S. Sakhare, M. Garcia Bardon, A. Sibaja-Hernandez, I. Ciofi, G. Eneman, A. Veloso, J. Ryckaert, P. Raghavan, A. Mercha, A. Mocuta, Z. Tökei, D. Verkest, P. Wambacq, K. De Meyer, and N. Collaert, "Vertical device architecture for 5nm and beyond: Device & Circuit implications", in *2015 Symposium on VLSI Technology (VLSI Technology)*, vol. 2015-Augus, IEEE, Jun. 2015, T26–T27.

5. M. Garcia Bardon, P. Schuddinck, P. Raghavan, D. Jang, D. Yakimets, A. Mercha, D. Verkest, and A. V.-Y. Thean, "Dimensioning for power and performance under 10nm: The limits of FinFETs scaling", in *2015 International Conference on IC Design & Technology (ICICDT)*, IEEE, Jun. 2015, pp. 1–4.

6. V. Huang, C. Pan, D. Yakimets, P. Raghavan, and A. Naeemi, "Device/system performance modeling of stacked lateral NWFET logic", in *2016 17th International Symposium on Quality Electronic Design (ISQED)*, IEEE, Mar. 2016, pp. 215–220.

7. P. Raghavan, M. Garcia Bardon, D. Jang, P. Schuddinck, D. Yakimets, J. Ryckaert, A. Mercha, N. Horiguchi, N. Collaert, A. Mocuta, D. Mocuta, Z. Tökei, D. Verkest, A. V.-Y. Thean, and A. Steegen, "Holistic device exploration for 7nm node", in *2015 IEEE Custom Integrated Circuits Conference (CICC)*, IEEE, Sep. 2015, pp. 1–5.

8. T. Huynh-Bao, S. Sakhare, J. Ryckaert, D. Yakimets, A. Mercha, D. Verkest, A. V.-Y. Thean, and P. Wambacq, "Design technology co-optimization for enabling 5nm gate-all-around nanowire 6T SRAM", in *2015 International Conference on IC Design & Technology (ICICDT)*, IEEE, Jun. 2015, pp. 1–4.

Curriculum Vitae

Dmitry Yakimets was born in Nizhny Tagil (Russia) in 1989. He received a B. Eng. degree in electronics design and manufacturing from Bauman Moscow State Technical University (Russia) in 2010, and a M. Sc. degree in nanotechnologies for ICT engineering from Grenoble Institute of Technology (France) in 2012. Within the framework of this Master program, in addition to his stay in France, Dmitry also studied in Polytechnic University of Turin (Italy) and Swiss Federal Institute of Technology in Lausanne (Switzerland). Since 2012, he has been working towards a PhD degree from KU Leuven at imec (Belgium). His research interests lay in the broad field of the design and technology co-optimization with the focus on a device modeling and benchmarking.

FACULTY OF ENGINEERING SCIENCE
DEPARTMENT OF ELECTRICAL ENGINEERING
ESAT / INSYS
Kasteelpark Arenberg 10 box 2440
B-3001 Leuven
dmitry@yakimets.xyz
http://yakimets.xyz